The Telecommunications and Data Acquisition Progress Report 42-91

July–September 1987

E. C. Posner
Editor

November 15, 1987

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The Telecommunications and Data Acquisition Progress Report 42-91

July–September 1987

E. C. Posner
Editor

November 15, 1987

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The research described in this publication was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Preface

This quarterly publication provides archival reports on developments in programs managed by JPL's Office of Telecommunications and Data Acquisition (TDA). In space communications, radio navigation, radio science, and ground-based radio astronomy, it reports on activities of the Deep Space Network (DSN) and its associated Ground Communications Facility (GCF) in planning, in supporting research and technology, in implementation, and in operations. Also included is TDA-funded activity at JPL on data and information systems and reimbursable DSN work performed for other space agencies through NASA. The preceding work is all performed for NASA's Office of Space Tracking and Data Systems (OSTDS).

In geodynamics, the publication reports on the application of radio interferometry at microwave frequencies for geodynamic measurements. In the search for extraterrestrial intelligence (SETI), it reports on implementation and operations for searching the microwave spectrum. The latter two programs are performed for NASA's Office of Space Science and Applications (OSSA).

Finally, tasks funded under the JPL Director's Discretionary Fund and the Caltech President's Fund which involve the TDA Office are included.

This and each succeeding issue of the TDA Progress Report will present material in some, but not necessarily all, of the following categories:

OSTDS Tasks:

- DSN Advanced Systems
  - Tracking and Ground-Based Navigation
  - Communications, Spacecraft-Ground
  - Station Control and System Technology
  - Network Data Processing and Productivity
- DSN Systems Implementation
  - Capabilities for Existing Projects
  - Capabilities for New Projects
  - New Initiatives
  - Network Upgrade and Sustaining
- DSN Operations
  - Network Operations and Operations Support
  - Mission Interface and Support
  - TDA Program Management and Analysis
- GCF Implementation and Operations
- Data and Information Systems

OSSA Tasks:

- Search for Extraterrestrial Intelligence
- Geodynamics
  - Geodetic Instrument Development
  - Geodynamic Science

Discretionary Funded Tasks

Contents

OSTDS TASKS
DSN Advanced Systems
TRACKING AND GROUND-BASED NAVIGATION

COMMUNICATIONS, SPACECRAFT–GROUND

STATION CONTROL AND SYSTEM TECHNOLOGY

NETWORK DATA PROCESSING AND PRODUCTIVITY

DSN Systems Implementation

CAPABILITIES FOR NEW PROJECTS

NETWORK UPGRADE AND SUSTAINING

OSSA TASKS
Search for Extraterrestrial Intelligence

A Demonstration of High Precision GPS Orbit Determination for Geodetic Applications

S. M. Lichten and J. S. Border
Tracking Systems and Applications Section

High precision orbit determination of Global Positioning System (GPS) satellites is a key requirement for GPS-based precise geodetic measurements and precise low-Earth orbiter tracking, which are currently being studied at the Jet Propulsion Laboratory (JPL). Different strategies for orbit determination have been explored at JPL with data from a 1985 GPS field experiment. The most successful strategy uses multi-day arcs for orbit determination and includes fine tuning of spacecraft solar pressure coefficients and station zenith tropospheric delays using the GPS data. Average rms orbit repeatability values for five of the GPS satellites are 1.0, 1.2, and 1.7 m in altitude, cross-track, and down-track components when two independent five-day fits are compared. Orbit predictions up to 24 hours outside the multi-day arcs agree within 4 m of independent solutions obtained with well-tracked satellites in the prediction interval. Baseline repeatability improves with multi-day as compared to single-day arc orbit solutions. When tropospheric delay fluctuations are modeled with process noise, significant additional improvement in baseline repeatability is achieved. For a 246 km baseline, with six-day arc solutions for GPS orbits, baseline repeatability is 2 parts in 10^8 (0.4–0.6 cm) for east, north, and length components and 8 parts in 10^8 for the vertical component. For 1314 and 1509 km baselines with the same orbits, baseline repeatability is 2 parts in 10^8 for the north components (2–3 cm) and 4 parts in 10^8 or better for east, length, and vertical components.

I. Introduction

The NAVSTAR Global Positioning System (GPS), when fully operational, will consist of a constellation of 18 satellites in 12-hour orbits designed to provide nearly continuous worldwide coverage for a variety of civil and military timing and positioning applications. For several years, the Jet Propulsion Laboratory (JPL) has been actively investigating GPS-based measurement systems with the goal of eventually demonstrating a sub-meter accuracy capability for GPS orbit determination. This capability will provide the basis for centimeter-level geodetic studies on a continental scale such as the NASA Geodynamics Program's Caribbean Initiative [1], [2] and decimeter-level positioning accuracy for low-Earth orbiters. A GPS-based measurement system will ultimately include GPS receivers situated at locations throughout the world to provide

accurate differential orbit determination for low-Earth orbiting satellites [3]–[5] such as TOPEX, the Earth Observing System, the Space Shuttle, and the Space Station, which will also be equipped with GPS receivers. Compared to other precise positioning systems such as very long baseline interferometry (VLBI), satellite laser ranging, and conventional ground tracking networks, GPS-based systems offer the potential for very high accuracy, an abundance of data, flexibility in the placement of ground receivers, and advantageous geometry provided by the high altitude and even distribution of the GPS satellites.

This article will discuss strategies for determining precise GPS orbits. Different aspects of orbit determination have been studied at JPL with 1985 experimental data processed through recently developed orbit and baseline estimation software. The 1985 GPS spring experiment took place between March 29 and April 5, 1985, representing the first of several GPS field tests conducted in 1985 and 1986. More than 20 institutions, including JPL, participated in the 1985 spring field test toward the goal of testing and evaluating GPS measurement techniques and equipment [1]. Data were acquired from seven developmental GPS satellites in orbit. TI-4100 GPS receivers [6] manufactured by Texas Instruments were placed at each of the ten ground sites in the continental United States. In addition, JPL SERIES-X receivers[1] were operated at the Mojave, California, and Owens Valley (OVRO), California, stations, and Air Force Geophysical Laboratory (AFGL) receivers, forerunners of the Macrometer II [7], were operated at the three POLARIS sites (Haystack, Massachusetts, Richmond, Florida, and Fort Davis, Texas). The POLARIS sites are collocated with VLBI radio telescopes used for geodetic and earth orientation studies. Water vapor radiometers (WVRs) were available at the Hat Creek, Mojave, and Owens Valley sites in California, while at other stations surface meteorological information was compiled during the experiment and used to correct the data for tropospheric delay. Table 1 and Fig. 1 summarize the experiment's configuration and the location of the ground receivers. Data were collected for about eight hours on each day of the experiment; however, because of the geometric constraints of the ground receiver configuration, GPS satellites were typically tracked for 2–4 hours before the receivers switched to new sets of satellites.

The GPS satellites transmit carrier signals at two L-band frequencies (1.22760 and 1.57542 GHz) which are modulated by a pseudo-random-noise code (P-code). All receivers in the field tests produce "carrier phase" observables from continuous tracking of the RF carriers. The TI-4100 receivers also produce a pseudorange observable by correlating the received modulated signal against a local copy of the code. The term pseudorange is used since this observable is a measure of the light travel time plus clock offsets at the satellite and the receiver. The carrier phase observable is analogous to "ambiguous range," since it measures range biased by an integer number of wavelengths. Both AFGL and SERIES-X receivers operate without knowledge of the P-code, although the SERIES-X receiver also generates a codeless pseudorange observable. Data from signals transmitted at the two frequencies are linearly combined to remove the dominant portion of the ionospheric delay, which varies inversely with the square of the frequency. Most of the results presented in this article are derived from the carrier phase data. Additional details on GPS signal structure and characteristics can be found in [8] and [9].

II. Data Processing

Data processing at JPL was conducted with the recently completed GIPSY (*G*PS *I*nferred *P*ositioning *SY*stem) software. GIPSY includes a comprehensive front end for data editing, phase connection, data compression, and atmospheric calibrations; PATH-VARY, a program which integrates the equations of motion and the variational equations to obtain nominal satellite trajectories and transition matrices for satellite states and dynamic model parameters; GPSOMC (*GPS O*bserved *M*inus *C*alculated), a module which calculates a very accurate model from best known nominal values and computes the pre-fit residuals and measurement partials; a U-D factorized batch sequential filter with process noise capabilities to perform parameter estimation and covariance analyses; and an output processor to display orbits, baselines, parameter solutions, and covariances. Clock modeling options include explicit single or double differencing, correlated process noise, and quadratic polynomials over specified data arcs. Additional details on specific modules in GIPSY can be found in [10].

Initially, GPS data processing, including parameter estimation, was performed separately for each day. Then multi-day arcs were formed, and longer runs were made covering up to six consecutive days. For some days, TI-4100 data were combined with the AFGL and SERIES-X data for orbit and baseline determination. However, since the different receiver types were in most cases collocated, there was not much to be gained (aside from an averaging effect on data noise) from processing all the data together—especially since with multi-day arcs, systematic effects as well as data noise become limiting error sources. Some of the TI-4100 data were used for determination of orbits and baselines independent of those based on the

[1] R. B. Crow, F. R. Bletzacker, R. J. Najarian, G. H. Purcell, J. I. Statman, and J. B. Thomas, "SERIES-X Final Engineering Report," JPL D-1476 (internal publication), Jet Propulsion Laboratory, Pasadena, California, 1984.

AFGL/SERIES-X network, and these separate solutions were compared to assess orbit and baseline repeatability.

III. Orbit Solution Strategy

The basic GPS orbit determination strategy includes simultaneous adjustment of the GPS orbits, station and satellite clock parameters, selected station locations, zenith tropospheric delays, range ambiguities (for carrier phase data), and solar pressure coefficients. There can be considerable variation in the way that these parameters are treated in the filter. In this section, various orbit determination strategies are described. In the next section, a comparison of orbit and baseline results using the different strategies is presented.

A. The Fiducial Concept

Fundamental to the GPS orbit determination performed at JPL is the fiducial network concept: a network of stations whose locations are accurately known from VLBI defines a self-consistent coordinate frame to which all GPS orbit and baseline solutions are referred. A detailed discussion of geophysical motivations for GPS geodesy, the selection of fiducial ground sites, and the validation of the fiducial concept with GPS measurements from the 1985 spring experiment can be found in [10]. Additional discussions of the role of GPS orbit determination in high precision geodesy can be found in [11] and [12]. An alternative to the fiducial approach, the free network approach, is discussed by Beutler *et al.* [13].

The results presented in this article are based on two basic fiducial network strategies. The first strategy has four fiducial ground stations, typically Haystack, Richmond, Fort Davis, and Owens Valley. This provides a well distributed set of reference points (Fig. 1). The second strategy has three fiducial points; the central Fort Davis station is adjusted along with the satellite orbits and other estimated parameters. Geometry provided by the second strategy is not as effective for defining the reference frame. However, in order to test baseline repeatability over long (>1000 km) distances, the second strategy was used for some of the solution sets. If Fort Davis or one of the other fiducial ground site locations were susceptible to systematic error or could be improved using the GPS data, the second strategy might be expected to produce better results.

B. Solar Radiation Pressure Model

The earliest GPS orbits determined at JPL showed that when fine tuning of the satellite solar pressure coefficients was not performed with simultaneous adjustment of station locations and GPS states, orbit and baseline repeatability was significantly degraded for data arcs longer than eight hours (single-day pass). If the solar radiation effects are left unmodeled, GPS position errors in down-track can increase to over 1 km after 1–2 weeks of integration of the equations of motion. Thus, even a small percentage error in one of the solar parameters can result in significant orbit errors. Solar radiation pressure is represented in the JPL software with the GPS Block I model [14], sometimes referred to as ROCK4. This model implicitly includes spacecraft component shapes, orientations, and masses and also accounts for inter-component shadowing. In ROCK4, it is assumed that the spacecraft remains perfectly oriented with respect to the sun. The spacecraft-centered coordinates defining the directions for solar radiation pressure accelerations (Fig. 2) have the Z-axis positive along the antenna directed toward the center of the Earth. The Y-axis is along the solar panel support beam, normal to the spacecraft–sun direction, and the X-axis is defined relative to the other two axes in the sense of a right-handed coordinate system. The force model represents the GPS spacecraft with 13 surfaces, each specified either as a flat surface or as a cylindrical surface. The flat surfaces are defined by length and width, while cylindrical surfaces are defined by radius of curvature, angle, and length. Each surface's reflectivity and specularity are represented by a number between 0 and 1 to account for varying absorption and diffusivity characteristics.

The solar radiation pressure results in a space vehicle acceleration

$$\ddot{\mathbf{r}} = P_1 \left[\frac{k^2}{\mathbf{r}_{ps}^2} (G_x a_x \hat{\epsilon}_x + G_z a_z \hat{\epsilon}_z) + G_y \hat{\epsilon}_y \right] \qquad (1)$$

where $\hat{\epsilon}_x$, $\hat{\epsilon}_y$, $\hat{\epsilon}_z$ are unit vectors for the spacecraft-centered coordinate system; the scalar P_1 is a shadow factor, 1 for direct sunlight, 0 for umbra, and between 0 and 1 for penumbra; $\mathbf{r}_{ps}$ is the spacecraft–sun vector; a_x and a_z are the space vehicle body fixed acceleration components determined by the ROCK4 model; k is the nominal Earth–sun distance (1 AU); G_x and G_z are solar pressure coefficient scaling factors; and G_y is a constant acceleration in the $\hat{\epsilon}_y$ direction, in km/sec^2, often referred to as the y-bias parameter.

The strategy adopted for fine tuning of the solar pressure parameters was to adjust G_x, G_y, G_z per satellite as constant parameters over the multi-day arc. Nominal solar radiation pressure parameters were taken from the precise post-fit ephemeris supplied by the Naval Surface Weapons Center (NSWC) [15]. The NSWC ephemeris was considered to be accurate in the WGS-72 reference frame to about 15–25 m (1 m in altitude and 15 m in each of the cross- and down-track components) and was determined from a two-week batch fit. The NSWC currently uses a more advanced software system known as the Multi-Satellite Filter/Smoother (MSF/S) [16] for generating reference GPS ephemerides. The *a priori*

uncertainties for the solar pressure coefficients for fitting at JPL from the March 1985 data were assumed to be 25 percent in x and z directions and 10^{-12} km/sec^2 for the y-bias. The NSWC fits in March and April 1985 vary from batch to batch typically by a few percent in the x and z coefficients and, for some satellites, by a significant fraction of 10^{-12} km/sec^2 for the y-bias. The JPL orbit strategy allowed somewhat more freedom in these solar pressure parameters in order to follow any shorter term variations which might be overlooked in the longer averaging period of the NSWC fits. In addition, the JPL orbits were determined with the x and z coefficients estimated as independent parameters. Nominally, these two parameters have the same value. This extra degree of freedom was left in the solution in order to absorb deficiencies in the solar pressure model and possibly to absorb other long term unmodeled accelerations. Some of the known limitations of the ROCK4 model, discussed by Fliegel *et al.* [17], could amount to orbit component errors of 4 m or more over a 14-day prediction interval. Allowing the JPL solar pressure parameters freedom to deviate from the nominal NSWC values was also motivated by the need to provide compensation for aliasing due to different parameter estimation strategies at JPL and NSWC. Finally, the geopotential model used in the NSWC fits (WGS-72 during March 1985) is slightly different from the GEM-L2 used in the JPL software. Because of these differences, NSWC and JPL force parameter solutions were expected to be slightly different.

The GPS x and z solar pressure parameter solutions obtained from a six-day arc covering March 31–April 5 were nearly all within 10 percent of the NSWC nominal values. The differences between the x and z coefficients were also generally less than 10 percent. The y-bias adjustments were somewhat larger, typically representing 25 percent changes from the nominal values. The adjustments for the GPS 6 solar pressure parameters were larger than those for the other satellites by a factor of about four. The reasons for this are not known, but these relatively large corrections persisted regardless of the manner in which the data were analyzed.

C. Clock Modeling Strategies

The GIPSY software used at JPL for GPS orbit determination offers a number of options for receiver and transmitter clock modeling. These options include clocks modeled as polynomials, explicit single and double differencing, and clock behavior modeled as process noise. Most of the results presented in this article have all clocks modeled as white noise; at each measurement time, the clocks are considered to be independent of their values at other times. The effect is very similar to clock elimination through double differencing [18], although the white noise treatment has the advantage of eliminating the introduction of correlations between measurements through explicit differencing.

D. Treatment of Tropospheric Delays

The troposphere is represented in GIPSY as a spherical shell which adds a delay to an incoming GPS signal:

$$\rho = \rho_{z_d} R_d(\theta) + \rho_{z_w} R_w(\theta) \tag{2}$$

where ρ_z is the tropospheric delay at zenith and R is an analytic mapping function developed by Lanyi [19] to map delays at zenith to the delay at elevation angle θ. The subscripts d and w refer to the dry and wet components of the tropospheric delay.

A priori values for zenith tropospheric delays were obtained from surface meteorology and, where available, from WVR measurements. The WVR measurements were converted to zenith delays using methods described by Robinson [20], and the surface meteorological measurements were converted to zenith delays using the Chao model [21].

The earliest GPS orbits determined at JPL relied on these *a priori* zenith delays for troposphere compensation. However, rms scatter was typically reduced by 50 percent or more when wet zenith delay corrections were estimated for each ground site in addition to the nominal calibrations. Various approaches to the troposphere estimation process were attempted, including not solving for tropospheric parameters at one or more WVR sites and varying the *a priori* uncertainty from 2 to 20 cm for the wet zenith delay parameter. For multi-day arcs, the zenith wet troposphere parameters were initially modeled with process noise in such a way that the troposphere values from one day to the next were independent but behaved as constants on any given day. Subsequently, in an effort to follow tropospheric fluctuations on time scales ranging from six minutes to several hours, the process noise time constraint was relaxed so that the tropospheric zenith delay could change slowly during the eight hour tracking period.

Two basic process noise models were used to model tropospheric fluctuations: colored noise and a random walk. The process noise formulation in the GIPSY filter is for first order exponentially correlated process noise, as described in detail by Bierman [22] and by Wu *et al.*[2] A brief summary of the formulation from those references is given here.

Let $p(t)$ be the value of a time-varying stochastic parameter and let $\omega(t)$ be a white process noise with zero mean value.

[2] S. C. Wu, W. I. Bertiger, J. S. Border, S. M. Lichten, R. F. Sunseri, B. G. Williams, P. J. Wolff, and J. T. Wu, "OASIS Mathematical Description," vol. 10, JPL D-3139 (internal publication), Jet Propulsion Laboratory, Pasadena, California, 1986.

The GIPSY filter and smoother are formulated in terms of discrete time intervals referred to as batches. At the end of the jth batch, the process noise parameters are updated:

$$p_{j+1} = m_j p_j + w_j \tag{3}$$

where $p_j = p(t_j)$ and $p_{j+1} = p(t_{j+1})$. The exponential multiplier, m_j, is defined as

$$m_j = \exp\left[-(t_{j+1} - t_j)/\tau\right] \tag{4}$$

where τ is the process noise time constant.

The discrete time-varying variance for a stochastic parameter is

$$\sigma^2_{p_{j+1}} = m_j^2 \sigma^2_{p_j} + q_{\text{dis}} \tag{5a}$$

$$= m_j^2 \sigma^2_{p_j} + (1 - m_j^2)\sigma^2_{ss} \tag{5b}$$

The steady-state sigma, σ_{ss}, is the noise level approached after the system has been operating undisturbed for a time much greater than τ. Together τ and σ_{ss} define the process noise variance in discrete form, q_{dis}.

The *random walk* is a special limiting case for Eq. (5). The random walk corresponds to $\tau \to \infty$. With a random walk, there is no steady state and σ_{ss} is not bounded, but q_{dis} is defined in the limiting sense

$$q_{\text{dis}} = \lim_{\tau \to \infty} \frac{\sigma^2_{ss}}{\tau} \tag{6}$$

Note that for a random walk, q_{dis} is related to the Allan variance $\sigma^2_A(\Delta t)$:

$$\sigma^2_A(\Delta t) = \frac{q_{\text{dis}}}{\Delta t} \tag{7}$$

Table 3 lists various process noise troposphere estimation strategies which were attempted in filtering the March 1985 GPS data. The second column of Table 3 shows the magnitude of the process noise for various models—either σ_{ss} for the correlated process noise cases or, for the random walk cases, the cumulative effect of q_{dis} [Eq. (6)] over one day.

IV. Orbit Determination Results

A. Coordinate Systems

Nominal orbits used to initialize the filter were supplied by the NSWC. In the spring of 1985, these orbits were computed in the WGS-72 coordinate system and were expected to be accurate to about 15–25 m. The fiducial stations used for the precise GPS orbit determination in this experiment are defined in the VLBI coordinate system, however, and thus a relatively large orbit adjustment was expected which would compensate for the difference (mostly a rotation in longitude) between the two coordinate systems. An accepted value for the coordinate rotation is 0.554 arc sec (2.6 μrad), although published values range from 0.5 to 0.8 arc sec [23]. A rotation of 0.554 arc sec corresponds (S. C. Wu *et al.*, see footnote 1) to approximately 15 m movement for a ground station and about 70 m at GPS altitude.

In order to empirically determine the coordinate system offset with the GPS data, a filter run was set up covering six days from March 31 to April 5. The only parameters estimated in this run were UT1–UTC, X and Y polar motion, and white process noise satellite and non-maser station clocks. Those station clocks which were using hydrogen masers as time standards were modeled in successive runs as linear and quadratic polynomials and as uncorrelated white process noise. The results were essentially independent of the clock model used for the masers. The solution for UT1–UTC was about 3.3 μrad, or approximately 0.68 arc sec. The X and Y polar motion solutions were at least an order of magnitude smaller. The difference between the JPL GPS result (0.68 arc sec) and the published value (0.554 arc sec) can be almost totally explained by the difference between the nominal earth orientation values supplied by the International Radio Interferometric Surveying Subcommission (IRIS) and those supplied by the Defense Mapping Agency (DMA). The IRIS values are used in GIPSY, while the DMA values were used by the NSWC for GPS orbit fitting. The difference between the DMA and IRIS UT1–UTC values for March 25, 1985, is 0.13 arc sec, which, when combined with the 0.554 arc sec assumed value for the WGS-72 to VLBI coordinate system rotation, is very close to the 0.68 arc sec determined for the JPL GPS solution.

Figure 3 shows two orbit fits for GPS 8 (we use the Navstar number to identify the satellites) from the six-day arc. Table 2 describes the basic orbit determination strategy used for these and most subsequent GPS solutions. Nearly all adjusted parameters were estimated with very large *a priori* uncertainties. The

data weights used for all solutions in this article were greater than the carrier phase intrinsic receiver noise of several millimeters in order to be consistent with post-fit rms scatter and in order to make $\chi^2_\nu \simeq 1$, where

$$\chi^2_\nu = \frac{\sum_{i=1}^{n} \left(\frac{z_i(\text{obs}) - z_i(\text{pred})}{\sigma_i} \right)^2}{n - m} \qquad (8)$$

In this expression, n and m are the number of measurements and degrees of freedom, z_i(obs) and z_i(pred) are the observed range and predicted post-fit range for the ith measurement, and σ_i is the measurement noise for the ith measurement. The data weights used were based on this criterion and correspond to measurement noise of about 0.9–1.5 cm, depending on the length of the data arc.

The solution plotted in Fig. 3(a) includes the coordinate system offset discussed above; most of the rotation shows up in the cross- and down-track orbit components. Figure 3(b) shows a solution from a filter run identical to the first except that a ~3 μrad coordinate rotation in UT1–UTC (longitude) was first included in the model before estimation of parameters. The large cross-track amplitude has been greatly reduced, since it is mostly due to the rotation of coordinate systems. The down-track run-off can be attributed to a slight overall altitude correction to the satellite ephemeris; because the orbital period changes with altitude, the down-track correction will slowly increase over time. An overall altitude adjustment of 70 cm, for example, would cause a run-off in the down-track component of about 7 m per orbit, roughly what is observed in Fig. 3(b). The NSWC nominal orbits are expected to be accurate to about 1 meter for the altitude component alone, so the results plotted in Fig. 3 are consistent.

B. Orbit Precision and Accuracy

Since the effective measurement noise was increased to be consistent with the post-fit rms scatter, the formal errors from the filter are one measure of the precision of the orbits. Figure 4 shows plots of formal orbit errors from a six-day arc solution and a single-day arc solution for GPS 8 and GPS 6 fit with carrier phase. GPS 8 was well tracked throughout the experiment, while GPS 6 was more sparsely tracked, with good geometry lasting only a few hours each day. The oscillatory nature of the error magnitude is due to tracking limitations, since the satellites were observed only from the continental United States each day. The single-day run included data from nine tracking sites, whereas the six-day run had only five to partially compensate for the greater quantity of data in the long arc. Figure 4 indicates that orbit precision of about 0.2–1.5 m is to be expected from multi-day arcs for well-tracked satellites, and about 0.5–3.0 m is to be expected for more sparsely tracked satellites. For one eight-hour pass (single-day arc), the formal errors are somewhat higher—1–3 m for GPS 8 and 2–8 m for GPS 6.

It is important to distinguish between orbit precision and accuracy. The formal orbit errors will in general underestimate orbit accuracy as a result of systematic effects which do not appear in post-fit scatter or which cannot be well compensated for simply by raising the effective measurement noise. A number of systematic errors which could grow slowly over time might not be apparent from post-fit scatter; such systematic effects could result from mismodeling of the earth's geopotential, from long-term spacecraft accelerations due to mismodeled solar radiation pressure or other unmodeled forces, or from earth orientation errors and fiducial station coordinate errors. Thus, comparisons between solutions derived from data sets in which different receivers and/or different data arcs were used may provide a better measure of orbit accuracy than would formal errors.

In order to better assess GPS orbit precision and accuracy, a set of single- and multi-day arcs was set up as shown in Fig. 5. Arcs A and B cover three days each and do not overlap in time. Arcs C and D cover five days each and do not have any data in common, although they overlap in time. Data on March 30 were also used to obtain GPS orbits completely independently of solutions obtained with the six-day arc covering March 31–April 5. These different sets of orbits were compared to measure orbit repeatability. In all the comparisons involving multi-day arcs, the solutions were compared over a "neutral" time interval outside the periods during which data were taken, so these comparisons provide a rather stringent test of the robustness of the orbits and models used to propagate and predict up to 24 hours outside the data arcs.

Figure 6(a) summarizes the GPS orbit repeatability for arcs A and B. These results show repeatability based on solutions in which the zenith troposphere at each site is adjusted as a constant bias independent from one day to the next. The overall average unweighted rms differences are 1.2, 2.9, and 2.6 m in altitude, cross-track, and down-track components for five satellites, averaged over a six hour period whose midpoint was about eight hours away from the nearest measurements. Figure 6(b) shows the corresponding GPS orbit repeatability for arcs C and D with constant zenith tropospheric delay parameters estimated. The overall average rms differences are reduced to 1.1, 1.5, and 2.4 m in altitude, cross-track, and down-track components. Figure 6(c) shows repeatability for arcs C and D with tropospheric zenith delay modeled with process noise as a random walk (first entry in Table 3). A noticeable improvement is apparent when Fig. 6(c) is compared to parts (a) and (b). With the process noise troposphere strategy, the average rms differences are further reduced to 1.0, 1.2, and 1.7 m in altitude, cross-track, and down-track components. Figure 7

shows sample orbit comparisons for GPS 8 and GPS 6 corresponding to the stochastic troposphere strategy of Fig. 6(c).

Figures 6 and 7 make a strong case for fine tuning the tropospheric delay parameters using process noise. Other troposphere process noise models showed similar improvement over the constant troposphere method, but the random walk model for tropospheric delay fluctuations gave the best orbit repeatability. Table 3 lists some of the troposphere strategies used and notes those which worked best as judged by orbit and baseline repeatability. Treuhaft and Lanyi [24] have developed a model for tropospheric delay fluctuations which behaves in a manner similar to a random walk over time scales of several minutes or longer and shows agreement with VLBI data. For shorter time scales, the Treuhaft and Lanyi model has characteristics of both a Markov process and a random walk. However, the tropospheric parameters estimated from the GPS data are residual delays due to limitations of the calibrations (from surface meteorology or WVRs) whose behavior is not easily modeled. We proceeded on the assumption that residual tropospheric delays after calibration behave in a manner similar to the tropospheric fluctuations themselves, although presumably with reduced amplitudes. All of the rms differences shown in Figs. 6 and 7 are consistent with the formal errors (1-σ) from the orbit fits. Figures 6 and 7 also seem to indicate that with essentially similar data and geometry, a five-day arc is preferable to a three-day arc for GPS orbit determination. However, it should also be noted that arcs A and B have no temporal overlap whatsoever; their repeatability is based on orbit prediction of up to 12 hours outside the data arcs and should be somewhat more sensitive to systematic errors which grow with time. The C-D repeatability, however, is based on interpolated orbits at times between five and eleven hours from measurements but still within the interleaved data arcs.

Figure 8 shows orbit repeatability using single day arcs. A single day typically included 6–7 hours of GPS data; GPS 8 was tracked for the entire data collection period, while the other spacecraft were tracked for periods ranging from one hour to four hours. Therefore, in contrast to the multi-day arcs, the amount of data and the geometry are very limited in a single day arc. The day shown in Fig. 8 is March 30. As typically happens during experiments of this kind, there were various equipment difficulties on March 30, including considerable data loss from the northeast which compromised viewing geometry. Although March 30 was not the worst day in terms of technical problems, for this experiment its data were definitely below average in quality as a consequence of data loss.

Figure 8(a) compares orbits from March 30 data for GPS 8 in which one orbit solution is from TI receivers and the other is from SERIES-X and AFGL receivers. Because one of the SERIES-X receivers (Owens Valley) was inoperative on March 30, the non-TI data set includes data from only four stations. Because of the data loss, the formal errors were relatively large when station clocks were modeled as white process noise. However, when the stations with maser time standards had their clocks modeled as polynomials, the formal errors decreased to a sufficient degree to make the comparison meaningful. The AFGL receivers at Richmond and Fort Davis and the SERIES-X receiver at Mojave were running on hydrogen maser clocks. The Haystack TI receiver used a crystal oscillator. The crystal clock fluctuations could be modeled as white process noise but not as a linear or quadratic polynomial. Figure 8(a) shows agreement at the 1 m level between the TI and non-TI orbits for March 30. It is instructive to see that even with a compromised data set, the stability of some of the station clock standards permitted clock modeling which improved the orbit for this satellite. Figure 8(b) compares the March 30 GPS 8 orbit with the GPS 8 orbit predicted back a full day from the six-day arc covering March 31–April 5; the orbits still agree at the 3 m level.

Figure 9 shows orbit repeatability with a one-day versus a multi-day arc for GPS 6, a satellite representative of those with somewhat sparse ground coverage. The single-day/multi-day comparison is not as good as with GPS 8, a well-tracked satellite (see Fig. 8). However, the multi-day/multi-day comparisons for GPS 6 match–and, for some components, even slightly surpass–repeatability for GPS 8 (see Figs. 6 and 7). Thus, without use of *a priori* knowledge of satellite trajectories, 1–2 m orbit repeatability is possible for both well tracked and sparsely tracked satellites with multi-day arcs, but only well-tracked satellites can achieve this level of repeatability with a single pass on one day for the limited configurations of this experiment. This is also reflected in the formal errors, which for sparsely tracked GPS were 5–10 m for the single day arc but decreased to 1–3 m with the multi-day arcs.

C. Baseline Repeatability

Baseline repeatability is often used to assess orbit quality in a geodetic context. Baseline repeatability tests have the advantage of not requiring any orbit propagation or prediction. A six-day arc was arranged to test baseline repeatability. Five ground receivers (Haystack, Richmond, Fort Davis, Mojave, and Owens Valley) were used in these solutions, with Fort Davis and Mojave station locations estimated each day independently. For each day of the multi-day arc, these station positions were reset with a large *a priori* uncertainty and were estimated anew with the data from that day. One set of smoothed orbits was generated based on all the data, including data from the two estimated stations. The rms scatters about the weighted means of the components of the Mojave–OVRO, Mojave–Fort Davis, and Fort Davis–OVRO baselines (246, 1314, and 1509 km) were calculated.

The baseline repeatability results are shown in Fig. 10. Figure 10(a) shows the Mojave–OVRO baseline, including a comparison of three different orbit estimation strategies: single-day pass solutions, multi-day arc solutions, and multi-day arcs with troposphere modeled as process noise. The multi-day arc solutions are more consistent than the single-day solutions for the 246 km baseline. The stochastic troposphere modeling strategy dramatically improves the 246 km baseline repeatability further; all components are consistent at the 0.5 cm level except for the vertical (2.0 cm). Parts (b) and (c), which give results for the longer baselines, show less dramatic improvement in all components than that seen in part (a) for the 246 km baseline with the troposphere process noise strategy. However, the components showing highest rms scatter still show moderate improvement. For the shorter baselines, it appears that fluctuating tropospheric delay can be a dominant systematic error source affecting repeatability if left unmodeled, but for the continental baselines, orbit errors and fiducial network errors may be more important. All baseline components over 246–1509 km show repeatability of 2–4 parts in 10^8 except for the vertical component for Mojave–OVRO (eight parts in 10^8).

It was noticed that Mojave–OVRO repeatability showed a slight but significant improvement when the Fort Davis position was adjusted as compared to solutions when Fort Davis was fixed as a fiducial station. This should not happen if all the fiducial station positions are known *a priori* to within a few centimeters. Therefore, it is possible that both the orbits and the baselines being estimated are affected by as yet undetermined errors in the Fort Davis or other fiducial reference site locations. In general, Fort Davis (either the TI receiver or the AFGL receiver) was not adjusted because its position was considered well determined from VLBI techniques and because it is geometrically advantageous to have a fixed reference station centrally located in the United States. For the baseline repeatability tests, however, Fort Davis was adjusted in order to provide a long (more than 1500 km) baseline result. The GPS data should be capable of isolating and improving the station(s) with errors in nominal receiver coordinates. When this is done, in the future it is likely that accuracy for both orbits and baselines will further improve.

D. Use of the Pseudorange

Previous covariance analyses [25] have indicated that either pseudorange with measurement noise of 10–20 cm (for five minute averaging periods) or carrier phase with measurement noise of about 1 cm should be capable of providing orbit accuracy of 2–3 m with eight hours of data. In the spring 1985 GPS experiment, the carrier phase noise for six minute measurement averages was consistently below 1 cm, and when used to determine GPS orbits, those orbits have repeatability values slightly better than the 2–3 m predicted from the covariance analysis. However, the pseudorange data scatter was found to be larger than the expected 20 cm. Figure 11 shows carrier phase and pseudorange rms post-fit scatter from a solution in which pseudorange and carrier phase were processed together. The carrier phase post-fit rms scatter was uniformly between 0.5 and 0.6 cm for nearly all the receivers; however, the pseudorange rms was much larger than expected and was highly variable, ranging from 68 cm to as high as 225 cm. Ground multipath has been identified as a likely source of much of the pseudorange scatter [26]. Satellite multipath may also have affected the data, but with an amplitude of less than 15 cm [27].

Tests at JPL showed that a significant fraction of the pseudorange scatter could be removed by subtracting a multipath signature from the data; the multipath signature was determined by isolating the signal patterns which repeat from one day to the next. Nevertheless, the residual scatter was still too large for the pseudorange to be useful in a high precision orbit and geodetic context. GPS orbits from a single-day pass of pseudorange had formal errors (as well as repeatability) of about 20–30 m, much higher than the 1–4 m (1–2 m for multi-day arcs) levels obtained with carrier phase.

Despite the high pseudorange data scatter, a limited demonstration was performed to show the potential benefit of pseudorange. With a one-day pass on March 30, the TI-4100 pseudorange and the carrier phase data were processed simultaneously. Each data type was weighted according to its rms scatter, so nearly all the solution strength came from the high quality carrier phase data. Orbit repeatability was measured by comparing these combined data type March 30 orbits with the carrier-phase-only orbits from a five-station, six-day arc covering March 31–April 5. For satellites with even moderately good coverage, only a slight improvement (if any) in orbit repeatability was seen with the March 30 combined data type orbits compared to March 30 orbits from carrier phase only. However, several of the satellites on March 30 had very short tracking periods and limited geometric coverage, adverse factors which were further exacerbated by data loss which occurred for that day. With only carrier phase from a single-day pass, these weakly tracked satellites had rather large (5–10 m) formal orbit errors. Single-day March 30 orbits for these satellites showed significant improvement when pseudorange data was added to the carrier phase measurements, despite the high scatter of the pseudorange data. Figure 12 shows the improvement for GPS 3. Note that the improvement is significant for the cross-track and down-track components; these components are believed to be weakest with carrier phase because of the necessity of solving for range ambiguity parameters, which relate the observed range change from carrier phase to the transmitter–receiver absolute range. This

weakness in cross-track and down-track is related to a weakness in baseline solutions for eastern components often seen in carrier-phase-only solutions.

A covariance analysis was performed to evaluate the potential of combining high quality pseudorange with carrier phase for orbit and baseline determination. The geometry of the 1985 spring experiment was used for this simulation, with seven ground stations, including three fiducial stations. Figure 13(a) shows how progressively better pseudorange can improve orbits when combined with carrier phase data, particularly in the cross-track and down-track components. Adding high quality pseudorange to the precise (but ambiguous) carrier phase has an effect similar to that of bias fixing [18], [28], [29], since the pseudorange provides an absolute range measurement which can effectively constrain the carrier phase ambiguity. Figure 13(b) shows that the east component of a baseline improves considerably with the addition of pseudorange to carrier phase data if the pseudorange measurement noise is less than 50 cm. As a point of comparison, parts (a) and (b) also show predicted orbit and baseline errors for carrier range, which is carrier phase data for which all range ambiguities have been resolved. Note that the mixed data types provide accuracies approaching that which would be attainable if carrier range were available. With a single eight-hour pass, the predicted orbit errors are at the 1 m level or better, and the east component of the baseline is determined to about 3 parts in 10^8. Since the 1985 spring experiment had a limited configuration of ground receivers and satellites, it is expected that performance will improve further when the full GPS constellation is available. Figure 13 includes an error contribution from *considered* parameters based on a 4 cm assumed uncertainty in each component of the fiducial station positions. The errors from considered parameters are based on the sensitivity of the orbits or baselines to unadjusted parameters; these sensitivities are calculated from the measurement partials and geometry [22].

If steps are taken in the future to reduce multipath (e.g., through antenna design and placement and through use of ground plane absorbers), the combined pseudorange and carrier phase data types could become a powerful tool in high precision orbit determination and geodetic studies. The accuracies which would be possible with carrier range could be approached with combinations of carrier phase and sufficiently precise pseudorange.

V. An Orbit Error Budget for GPS

Formal orbit errors (Fig. 4) are based on the assigned data weights and on the *a priori* covariance for the estimated parameters. Since all the estimated parameters (except perhaps the tropospheric zenith delays) were basically unconstrained initially in the solutions, the data weights essentially determine what the formal orbit errors are. The data weights were set to be equal to the measurement noise scaled by a factor in accordance with Eq. (8) so that the weights were consistent with the post-fit scatter and with the number of measurements and estimated parameters. Typically, this scaling factor was between 1.0 and 1.5, which indicates that some residual systematic errors could be affecting the orbit solutions.

A consider covariance analysis was performed to examine the sensitivity of the orbit solutions to unadjusted parameters and to various potential systematic errors. As discussed above, the errors from considered parameters are combined with the formal errors computed in the filter to produce a total error covariance which includes the effects of likely systematic error sources whose effects are not compensated for in the basic solution strategy. Models for these unestimated parameters are used to calculate their effect on the orbits. Systematic errors included in this consider analysis were fiducial station location errors, errors in Earth orientation parameters, and uncertainty in the location of the Earth's geocenter relative to the fiducial stations. The consider analysis covered a six-day arc with five ground stations, including fiducials at Owens Valley, Richmond, and Haystack. Table 4 describes the assumptions for the analysis. The uncertainties in the earth orientation parameters specifying UT1-UTC and X and Y polar motion are based on the results of Spieth *et al.* [30] and Steppe *et al.* [31]. These uncertainties apply to earth orientation parameters which might be available several weeks after a GPS experiment; for real-time applications, uncertainties in the earth orientation parameters would be expected to be several times higher. Figure 14 shows the orbit error breakdown for GPS 8. Note that the expected root-sum-square (RSS) orbit errors are close to the 1–2 m orbit repeatabilities observed with multi-day arc (Fig. 6[c]). The two largest error categories are the computed (formal) error and the systematic errors from uncertainties in the fiducial station locations.

The computed error could be reduced by taking more measurements, lengthening the arc, including *a priori* information, or adding pseudorange. The number of measurements could be increased by the addition of more ground receivers, but as more GPS satellites are launched in the future, the amount of data will increase naturally. More *a priori* information could be included by using satellite and station clock models, WVR tropospheric calibrations at more sites, and better nominal orbits. Perhaps the most dramatic decrease in the computed error could be achieved if bias fixing were possible over long baselines [29]. With bias fixing, the east baseline component (corresponding to cross- and down-track orbit components) computed errors would decrease by at least a

factor of two. Even without bias fixing, the addition of pseudorange would similarly reduce the computed error, since it tightly constrains the carrier phase ambiguity parameters. With some types of receivers, pseudorange is available in addition to the carrier phase, but multipath must be reduced considerably before it is of sufficient quality to obtain sub-meter orbits in one pass. Lengthening the data arc beyond six days is feasible, but it might be more costly than some of these other remedies and might increase sensitivity to systematic errors from unmodeled accelerations.

Even if the computed contribution were reduced by a factor of two, the RSS orbit errors would still be over 1 m in cross-track and down-track if fiducial station position errors are 4 cm. Improvement in the fiducial station network requires painstaking refinement in the ties between the geodetic monuments and the GPS and VLBI antennas. Because it appears to be one of the limiting error sources, fiducial station accuracy is an area of intensive study at JPL.

The remaining systematic error sources in Fig. 14 are the earth orientation parameters (UT1–UTC, X and Y polar motion) and the location of the geocenter. These contribute relatively little to the total orbit error under the assumptions of the consider analysis but could become significant in the future as sub-meter GPS accuracy is approached. Note that although the earth orientation errors have a small but noticeable contribution to the GPS orbit error budget, the effect of these errors on differential measurements such as baselines is insignificant compared to other error sources [10]. The UT1–UTC error contribution in Fig. 14 applies only to the inertial reference frame (J2000); in an earth-fixed frame, the UT1–UTC error is eliminated in the transformation as long as the same value used in the orbit solutions is used to rotate to the earth-fixed frame. Most geodetic GPS applications, such as baseline determination, are based in an earth-fixed reference frame; thus, the UT1–UTC error would have essentially no effect, although it would introduce a bias in the GPS orbits.

Other systematic orbit errors result from unmodeled forces and accelerations which can affect the spacecraft trajectories. This category of errors includes spacecraft accelerations due to gravity mismodeling, gas leaks, solar radiation pressure mismodeling, and atmospheric drag. We believe that unmodeled accelerations of this type are small for GPS compared to other systematic effects, at least for arcs of one to several days. Preliminary consider analysis [25] shows very small perturbations (less than 5 cm down-track errors) from gravity field mismodeling for GPS orbits determined from carrier phase. In a future study, the analysis of gravity errors will be extended to cover longer arcs and to include the gravity field covariance matrix to model errors in the gravity coefficients. Since three solar radiation pressure coefficients were estimated for each satellite, we believe that our model has sufficient freedom to compensate for the solar radiation pressure forces as well as for some other minor spacecraft accelerations. When longer data arcs are available, these effects will be studied in more detail. Finally, the covariance studies found atmospheric drag at GPS altitudes to be insignificant, at least for data arcs of up to several weeks.

VI. Summary and Conclusions

Data from the spring 1985 GPS field test have been processed, and precise GPS orbits have been determined. With carrier phase data spanning five days, orbit repeatability is 1–2 m for each component, averaged for five satellites over a six hour period during which no data were taken. Baseline repeatability over the one-week period for a baseline of 246 km is about 2 parts in 10^8 (0.4–0.6 cm) for east, north, and length components; vertical repeatability is several times worse (8 parts in 10^8) for this baseline. Repeatability for two baselines of 1314 and 1509 km is 2–4 parts in 10^8 (3–5 cm) for all components.

Several refinements to the orbit determination strategies were found to be crucial to achieving these levels of repeatability and accuracy. These include fine tuning the GPS solar radiation coefficients and ground station zenith tropospheric delays. The time-varying behavior of the troposphere was modeled with process noise for the best results. Multi-day arcs of three to six days provided better orbits and baselines than the eight-hour arcs from single-day passes. A limited demonstration was able to show the potential for further orbit and baseline accuracy improvement with combined pseudorange and carrier phase data.

References

[1] J. M. Davidson, C. L. Thornton, C. J. Vegos, L. E. Young, and T. P. Yunck, "The March 1985 Demonstration of the Fiducial Network Concept for GPS Geodesy: A Preliminary Report," in *Proceedings of the First International Symposium on Precise Positioning with the GPS*, pp. 603–611, 1985.

[2] T. H. Dixon, M. P. Golombek, and C. L. Thornton, "Constraints on Pacific Plate Kinematics and Dynamics with Global Positioning System Measurements," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-23, pp. 491–501, July 1985.

[3] T. P. Yunck, W. G. Melbourne, and C. L. Thornton, "GPS-Based Satellite Tracking System for Precise Positioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-23, pp. 450–457, 1985.

[4] T. P. Yunck, S. C. Wu, and S. M. Lichten, "A GPS Measurement System for Precise Satellite Tracking and Geodesy," *J. Astronautical Sci.*, vol. 33, pp. 367–380, 1985.

[5] S. M. Lichten, S. C. Wu, J. T. Wu, and T. P. Yunck, "Precise Positioning Capabilities for TOPEX Using Differential GPS," *AAS Paper 85-401*, AAS/AIAA Astrodynamics Specialist Conference, Vail, Colorado, August 1985.

[6] D. J. Henson, E. A. Collier, and K. R. Schneider, "Geodetic Applications of the Texas Instruments TI 4100 GPS Navigator," in *Proceedings of the First International Symposium on Precise Positioning with the GPS*, pp. 191–200, 1985.

[7] J. W. Ladd and C. C. Counselman III, "The Macrometer II Dual-Band Interferometric Surveyor," in *Proceedings of the First International Symposium on Precise Positioning with the GPS*, pp. 175–180, 1985.

[8] R. J. Milliken and C. J. Zoller, "Principle of Operation of NAVSTAR and System Characteristics," *Navigation*, vol. 25, pp. 95–106, 1978.

[9] E. H. Martin, "GPS User Equipment Error Models," *Navigation*, vol. 25, pp. 201–210, 1978.

[10] J. M. Davidson, C. L. Thornton, S. A. Stephens, G. Blewitt, S. M. Lichten, O. J. Sovers, P. M. Kroger, L. L. Skrumeda, J. S. Border, R. E. Neilan, C. J. Vegos, B. G. Williams, J. T. Freymueller, T. H. Dixon, and W. G. Melbourne, " The Spring 1985 High Precision Baseline Test of the JPL GPS-Based Geodetic System: A Final Report," to be published as a JPL publication, 1987.

[11] R. I. Abbot, Y. Bock, C. C. Counselman III, R. W. King, S. A. Gourevitch, and B. J. Rosen, "Interferometric Determination of GPS Satellite Orbits," in *Proceedings of the First International Symposium on Precise Positioning with the GPS*, pp. 63–72, 1985.

[12] G. Beutler, W. Gurtner, I. Bauersima, and R. Langley, "Modeling and Estimating the Orbits of GPS Satellites," in *Proceedings of the First International Symposium on Precise Positioning with the GPS*, pp. 99–111, 1985.

[13] G. Beutler, W. Gurtner, M. Rothacher, and I. B. Schildknecht, "Evaluation of the March 1985 High Precision Baseline (HPBL) Test: Fiducial Point Concept versus Free Network Solutions," *EOS Transactions, American Geophysical Union*, vol. 67, p. 911, Nov. 4, 1986.

[14] W. W. Porter, "Solar Force-Torque Model for the GPS Space Vehicle System," Rockwell International Space Division, Downey, California, February 18, 1976.

[15] S. S. Russel and J. H. Schaibly, "Control Segment and User Performance," *Navigation*, vol. 25, pp. 74–80, 1978.

[16] E. R. Swift, "NSWC's GPS Orbit/Clock Determination System," in *Proceedings of the First International Symposium on Precise Positioning with the GPS*, pp. 51–62, 1985.

[17] H. F. Fliegel, W. A. Feess, W. C. Layton, and N. W. Rhodus, "The GPS Radiation Force Model," in *Proceedings of the First International Symposium on Precise Positioning with the GPS*, pp. 113–119, 1985.

[18] Y. Bock, S. A. Gourevitch, C. C. Counselman III, R. W. King, and R. I. Abbot, "Interferometric Analysis of GPS Phase Observations," *Manuscripta Geodaetica*, vol. 11, pp. 282–288, 1986.

[19] G. Lanyi, "Troposphere Calibration in Radio Interferometry," in *Proceedings of the International Symposium on Space Techniques for Geodynamics*, p. 184, 1984.

[20] S. E. Robinson, "A New Algorithm for Microwave Delay Estimation from Water Vapor Radiometer Data," *TDA Progress Report 42-87*, vol. July–September 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 149–157, November 15, 1986.

[21] C. C. Chao, *A New Method to Predict Wet Zenith Range Correction from Surface Measurements*, JPL Technical Report 32-1526, Jet Propulsion Laboratory, Pasadena, California, 1973.

[22] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, New York: Academic Press, 1977.

[23] L. D. Hothem, T. Vincenty, and R. E. Moose, "Relationship Between Doppler and Other Advanced Geodetic System Measurements Based on Global Data," in *Proceedings of the Third International Geodetic Symposium on Satellite Doppler Positioning*, vol. 1, pp. 109–128, 1982.

[24] R. N. Treuhaft and G. E. Lanyi, "The Effect of the Dynamic Wet Troposphere on Radio Interferometric Measurements," *Radio Science*, vol. 22, pp. 251–265, 1987.

[25] W. I. Bertiger, S. C. Wu, J. S. Border, S. M. Lichten, B. G. Williams, and J. T. Wu, "High Precision GPS Orbit Determination Using March 1985 Demonstration Data," AIAA Paper 86-0089, AIAA 24th Aerospace Sciences Meeting, Reno, Nevada, January 6–9, 1986.

[26] A. E. Evans, "Comparison of GPS Pseudorange and Biased Doppler Range Measurements to Demonstrate Signal Multipath Effects," in *Proceedings of the Fourth International Geodetic Symposium on Satellite Positioning*, pp. 573–587, 1986.

[27] R. E. Neilan, "An Experimental Investigation of the Effects of GPS Satellite Multipath," master's thesis, University of Wisconsin, Madison, Wisconsin, 1986.

[28] Y. Bock, C. C. Counselman III, S. A. Gourevitch, and R. W. King, "Establishment of Three-dimensional Control by Interferometry with the Global Positioning System," *Journal of Geophysical Research*, vol. 90, pp. 7689–7703, 1985.

[29] P. L. Bender and D. R. Larden, "GPS Carrier Phase Ambiguity Resolution Over Long Baselines," in *Proceedings of the First International Symposium on Precise Positioning with the GPS*, pp. 357–361, 1985.

[30] M. A. Spieth, T. M. Eubanks, and J. A. Steppe, "Intercomparison of UT1 Measurements During the MERIT Campaign Period," in *Proceedings of the International Conference on Earth Rotation and the Terrestrial Reference Frame, Part II*, vol. 2, pp. 609–621, 1985.

[31] J. A. Steppe, T. M. Eubanks, and M. A. Spieth, "Intercomparison of Polar Motion Measurements During the MERIT Period," in *Proceedings of the International Conference on Earth Rotation and the Terrestrial Reference Frame, Part II*, vol. 2, pp. 622–636, 1985.

Table 1. Receiver deployment

Location	Receiver	WVR
Austin, Texas	TI-4100	–
Dahlgren, Virginia	TI-4100	–
Fort Davis, Texas	TI-4100 AFGL	–
Hat Creek, California	TI-4100	Yes
Haystack, Massachusetts	TI-4100 AFGL	–
Mammoth Lakes, California	TI-4100	–
Mojave, California	TI-4100 JPL SERIES-X	Yes
Owens Valley, California	TI-4100 JPL SERIES-X	Yes
Point Mugu, California	TI-4100	–
Richmond, Florida	TI-4100 AFGL	–

Table 2. Parameter estimation strategy

Parameter	A priori σ
Satellite positions	(20, 20, 20) km
Satellite velocities	(2, 2, 2) m/s
Non-fiducial station locations	(1, 1, 1) km
White noise clocks	0.1 sec
Polynomial clocks	0.1 sec bias 10^{-7} sec/sec rate 10^{-13} sec/sec^2 accel
Carrier phase bias	10 sec
Zenith wet tropospheric delay	20 cm (no WVR) 3 cm (WVR)
Solar radiation pressure	
X, Z, Y-bias coefficients	25%, 25%, 100%
Data weights: 1–2 cm	(white noise clocks all arcs)
2 cm	(polynomial clocks 1-day arc)
Data interval: 1 carrier ϕ meas/360 sec	

Table 3. Stochastic troposphere estimation strategies

Strategy	Process noise level	τ	σ_p a priori
1. Random walk	10 cm except Fort Davis/OVRO 3 cm Fort Davis/OVRO	∞	20 cm (SM) 3 cm (WVR)
2. Random walk	10 cm except Fort Davis 3 cm Fort Davis	∞	20 cm (SM) 3 cm (WVR)
3. Random walk	10 cm except Fort Davis 3 cm Fort Davis	∞	20 cm
4. Random walk	10 cm	∞	20 cm (SM) 3 cm (WVR)
5. Colored noise	σ_{ss} = 5 cm	12 hrs (SM) 24 hrs (WVR)	20 cm
6. Colored noise	σ_{ss} = 5 cm	12 hrs	20 cm
7. Colored noise	σ_{ss} = 10 cm	96 hrs	20 cm
8. Random walk	7.5 cm	∞	20 cm

ρ: zenith tropospheric delay
σ_ρ: uncertainty in ρ
σ_{ss}: steady state colored noise sigma Eq. (5)
Process noise level:
For random walks: cumulative effect of process noise [q_{dis} in Eq. (6)] on σ_ρ over 24 hrs
For colored noise: σ_{ss}
τ: colored noise exponential correlation time constant
WVR: ρ calibrated with water vapor radiometer measurements
SM: ρ calibrated with surface meteorology measurements

Process noise models listed in order of performance based on baseline/orbit repeatability.
Models 1–4 performed significantly better than models 5–8 or models with constant ρ.

Table 4. Consider analysis assumptions

Consider parameter	Consider σ
Fiducial coordinates	(4, 4, 4) cm
UT1–UTC	0.2 msec
X-pole	2 masec
Y-pole	2 masec
Geocenter coordinates	(10, 10, 10) cm

Five ground stations, six-day tracking arc
Fiducials: Owens Valley, Richmond, Haystack
Station and satellite clock model: white noise
Troposphere model: random walk
Carrier ϕ data interval: 1 meas/360 sec
Data weight: 1.5 cm

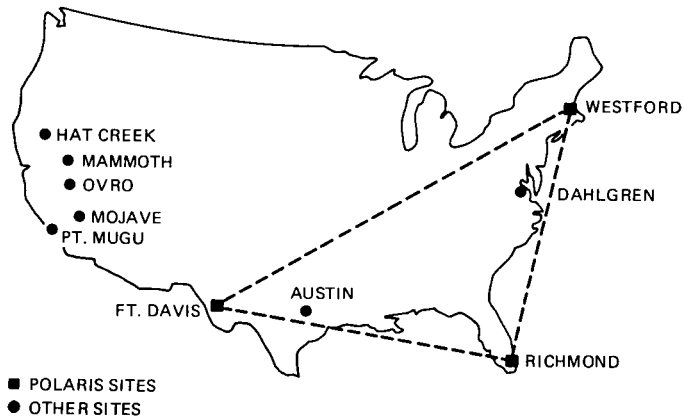


Fig. 1. Locations of GPS receivers during the spring 1985 GPS experiment

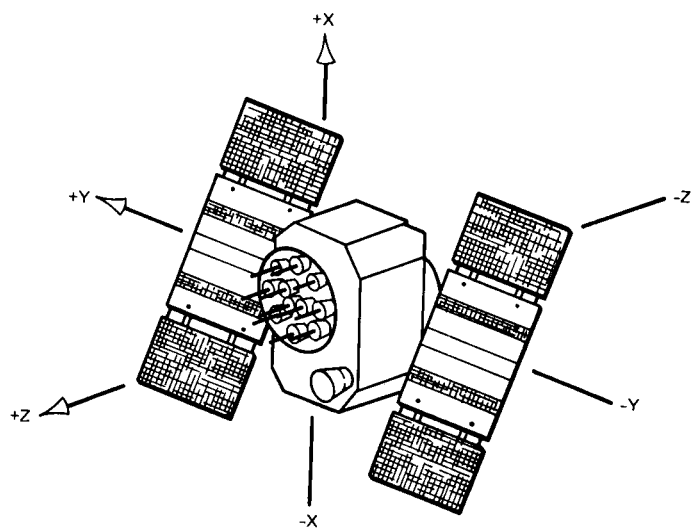


Fig. 2. Local coordinate system used to define GPS solar pressure coefficients in the ROCK4 model

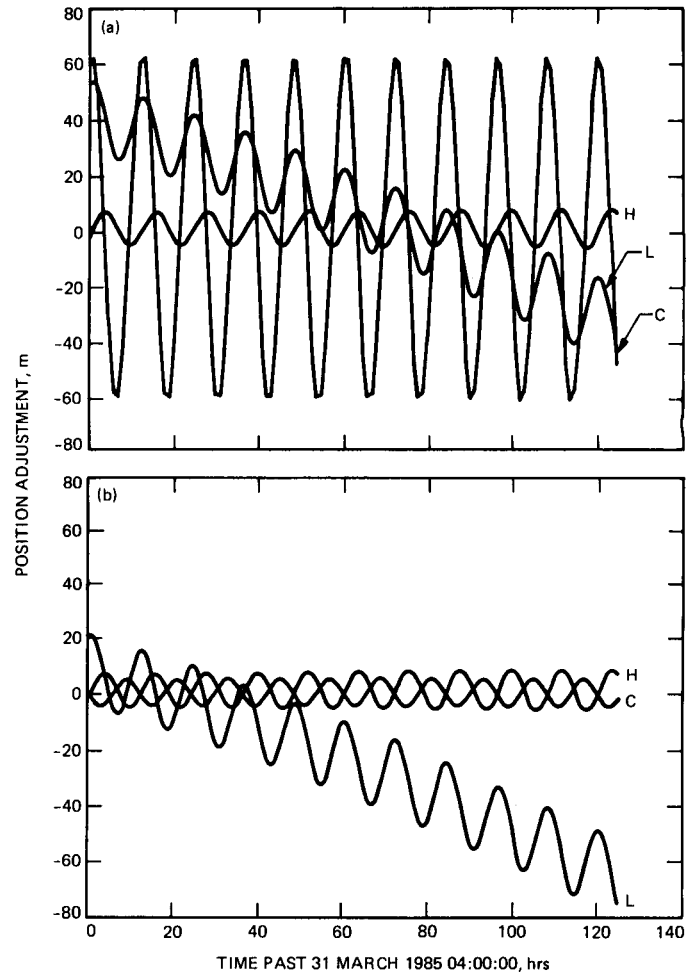


Fig. 3. Altitude (H), cross-track (C), and down-track (L) orbit adjustments for GPS 8 from a six-day arc: (a) with large orbit corrections, mostly due to coordinate system offset; (b) with ~ 3 μrad coordinate system rotation removed before orbit determination

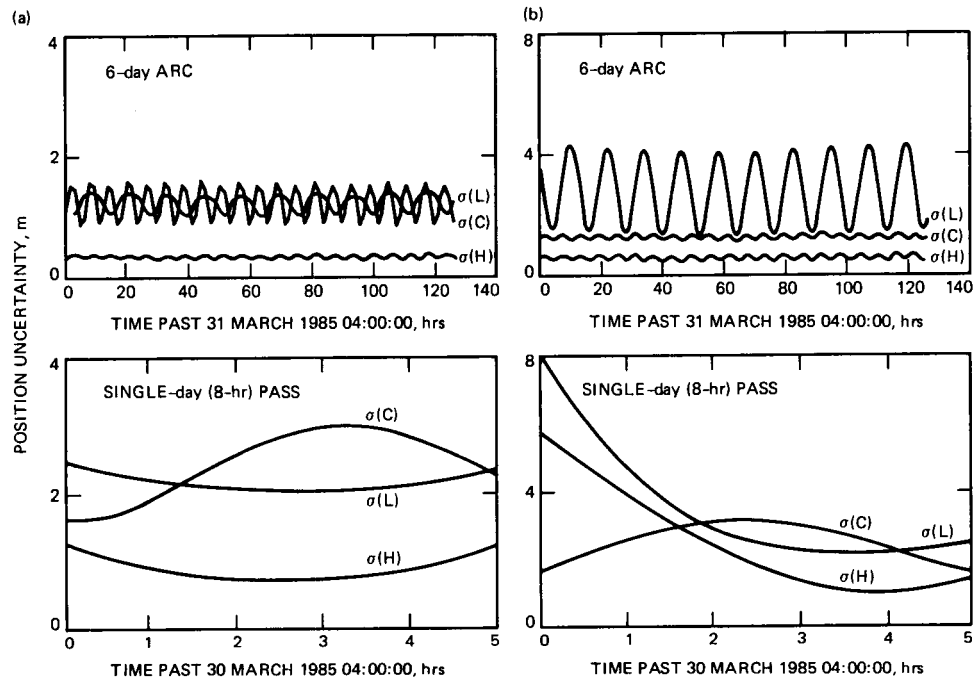


Fig. 4. Formal orbit uncertainties for satellites tracked from a six-day arc with five ground stations and from a single-day (eight-hour) pass with nine ground stations: (a) a well-tracked satellite (GPS 8); (b) a sparsely tracked satellite (GPS 6)

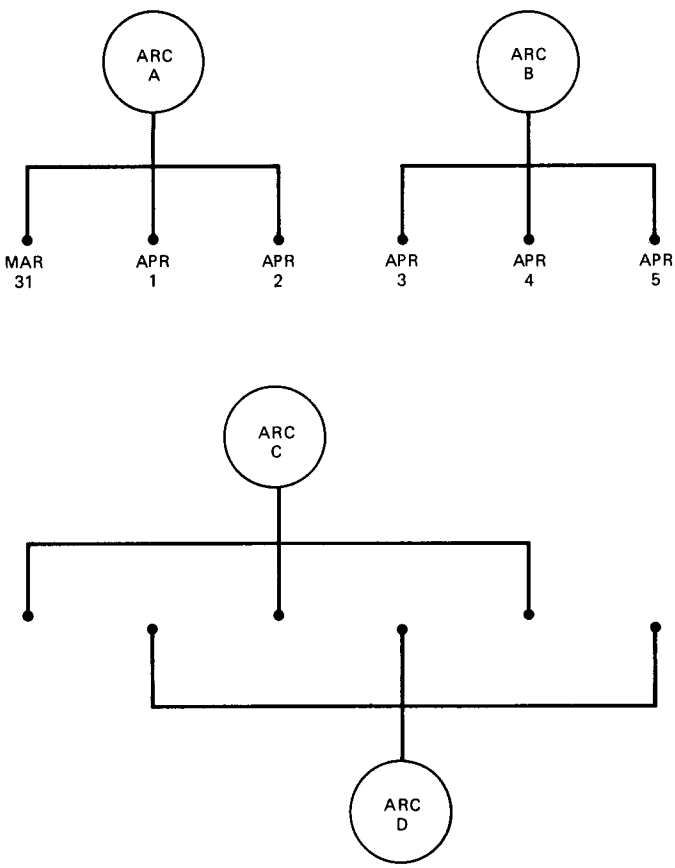


Fig. 5. Arcs A, B, C, and D used for orbit repeatability studies

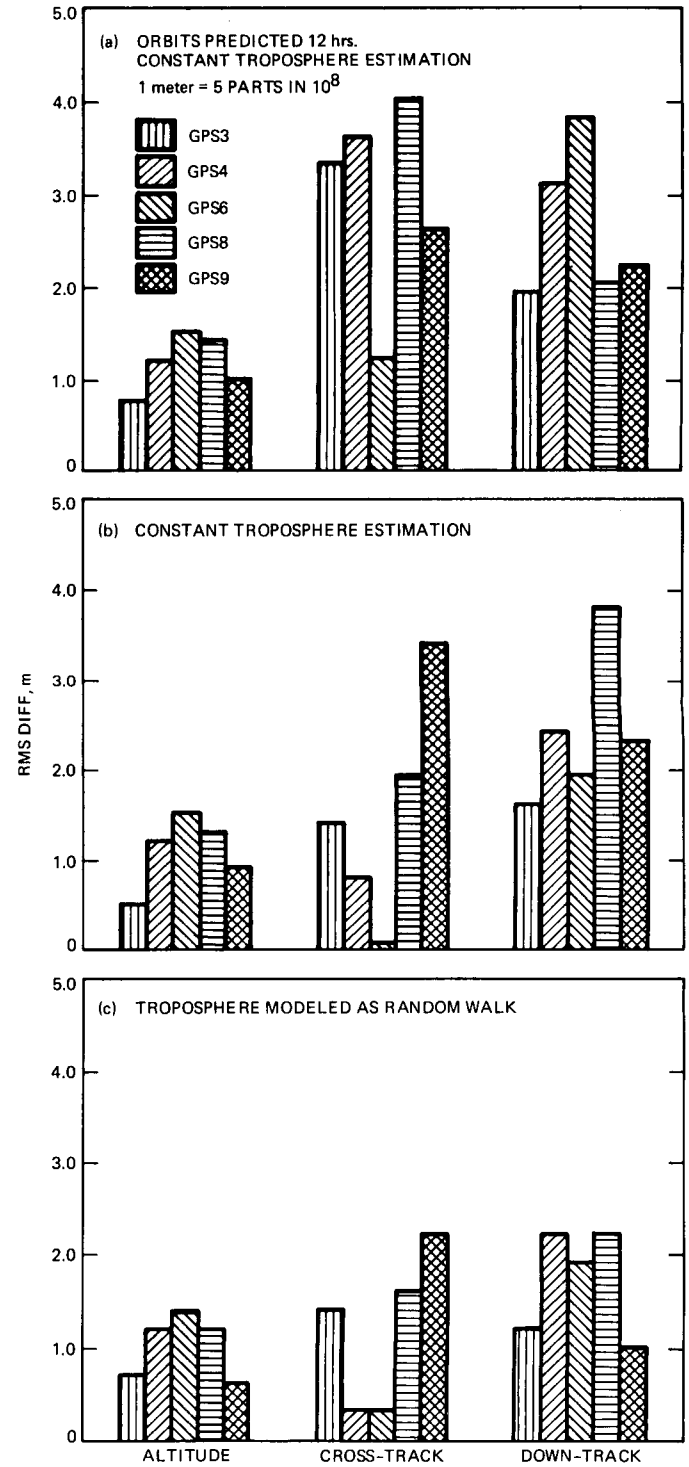


Fig. 6. Orbit repeatability for five satellites: (a) using arcs A and B with constant zenith wet tropospheric delays estimated at each station; (b) using arcs C and D with constant zenith wet tropospheric delays estimated at each station; and (c) using arcs C and D with random walk process noise models for the zenith wet tropospheric delays estimated at each station

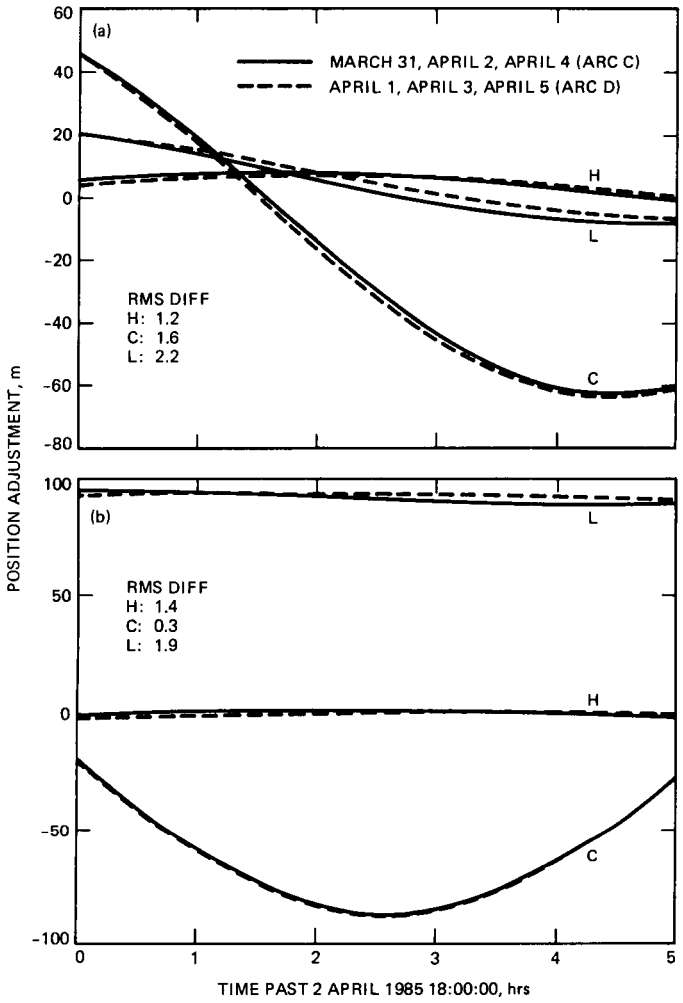


Fig. 7. Comparison of orbit solutions for (a) GPS 8 and (b) GPS 6 mapped to a six-hour period during which no observations were taken; these solutions correspond to results plotted in Fig. 6(c)

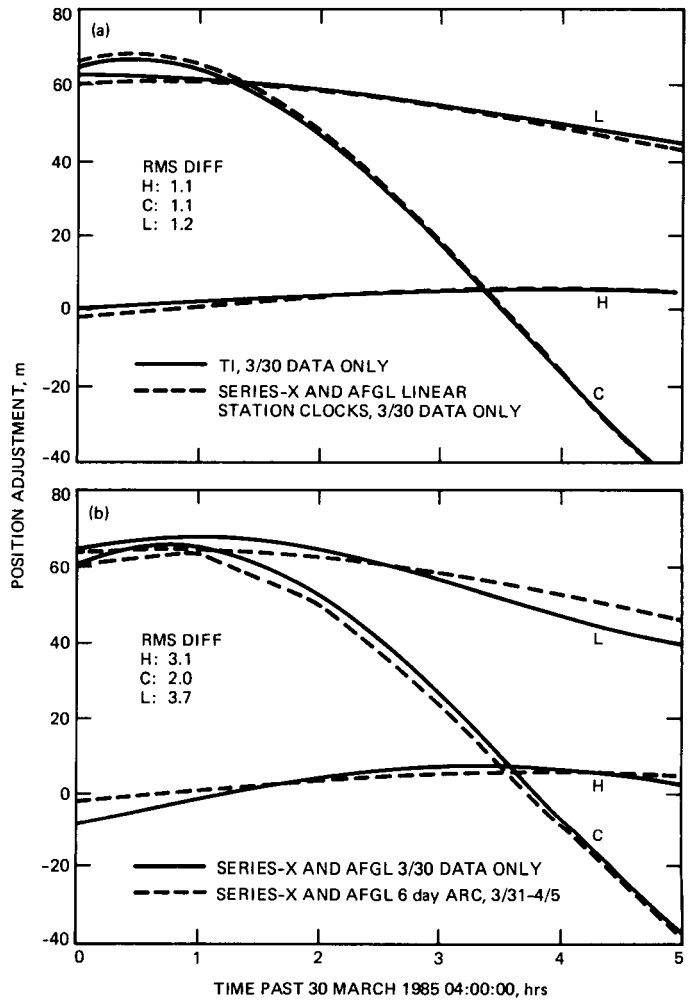


Fig. 8. Orbit repeatability for GPS 8 with (a) two single-day arcs and (b) a single-day arc and multi-day arc solutions

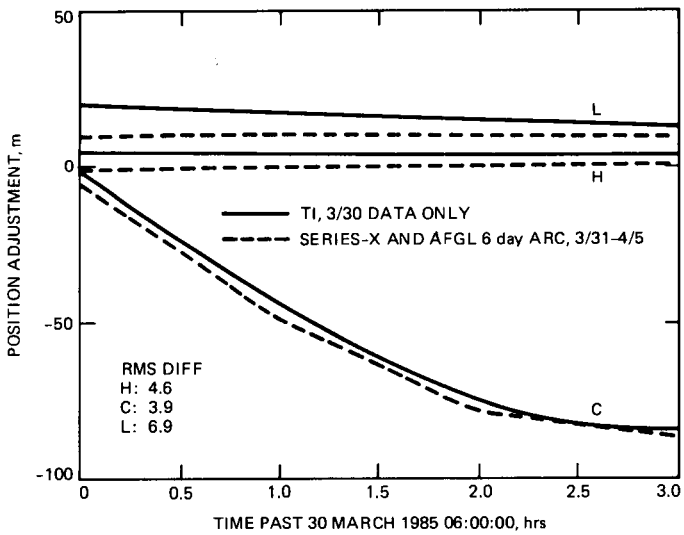


Fig. 9. Comparison of single-day and multi-day orbit solutions for GPS 6

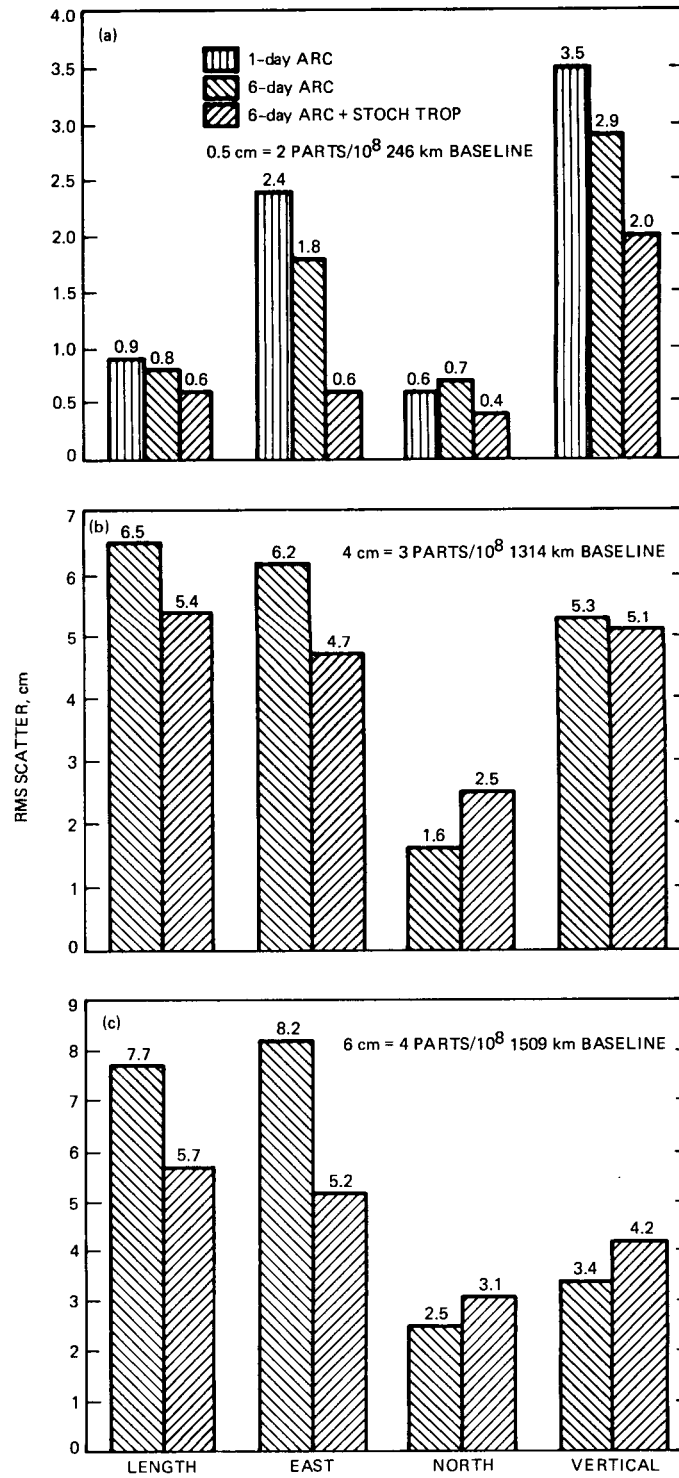


Fig. 10. Baseline repeatability for (a) Mojave-OVRO; (b) Mojave-Fort Davis; and (c) OVRO-Fort Davis

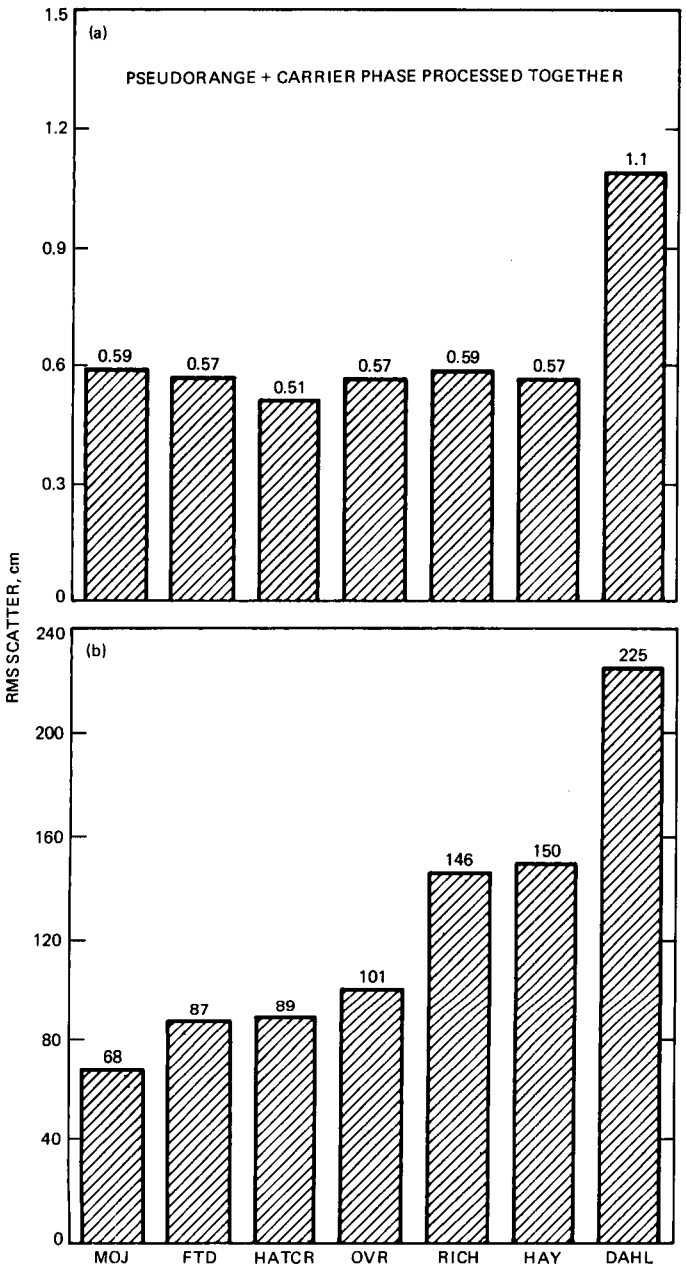


Fig. 11. Post-fit rms scatter for (a) carrier phase and (b) pseudorange from a solution in which both data types were processed together; post-fit scatter is shown for seven TI-4100 receivers

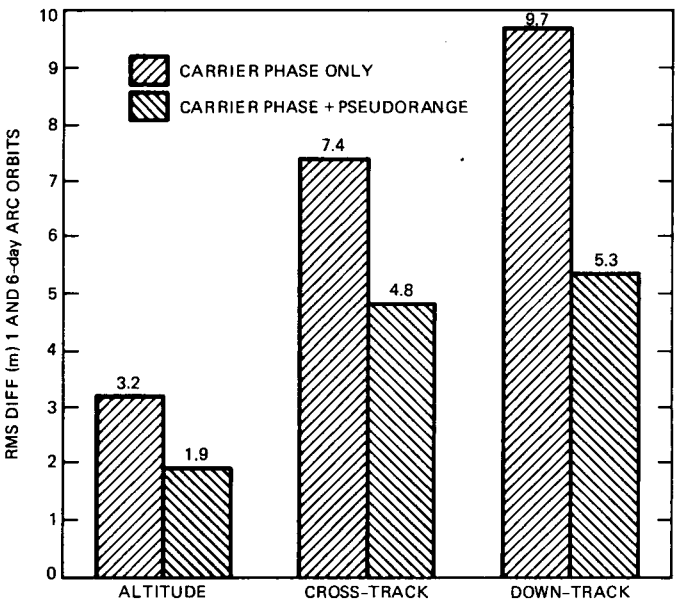


Fig. 12. Improvement in orbit repeatability resulting from processing pseudorange data together with the integrated doppler from carrier phase for a single-day pass; the single-day pass solutions are compared with a six-day, carrier-phase-only solution

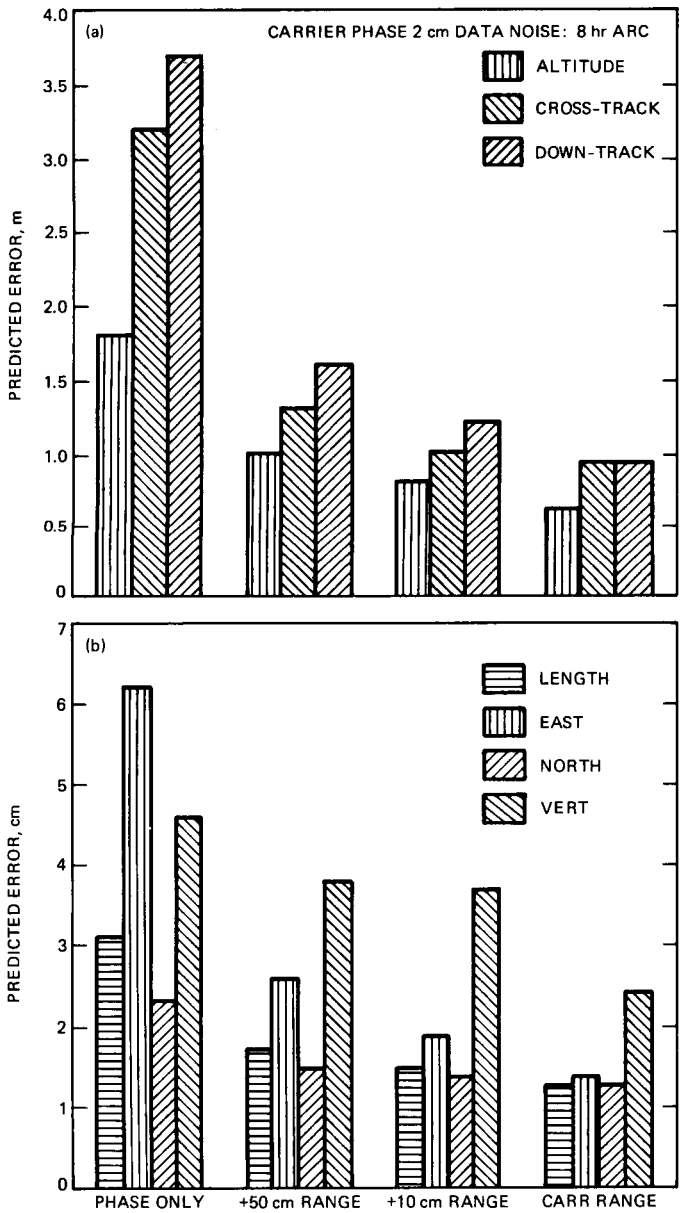


Fig. 13. Predicted reduction of orbit and baseline errors from covariance analysis when carrier phase and pseudorange data are processed simultaneously, assuming 4 cm uncertainty in fiducial station coordinates: (a) predicted orbit accuracy; (b) predicted baseline determination accuracy

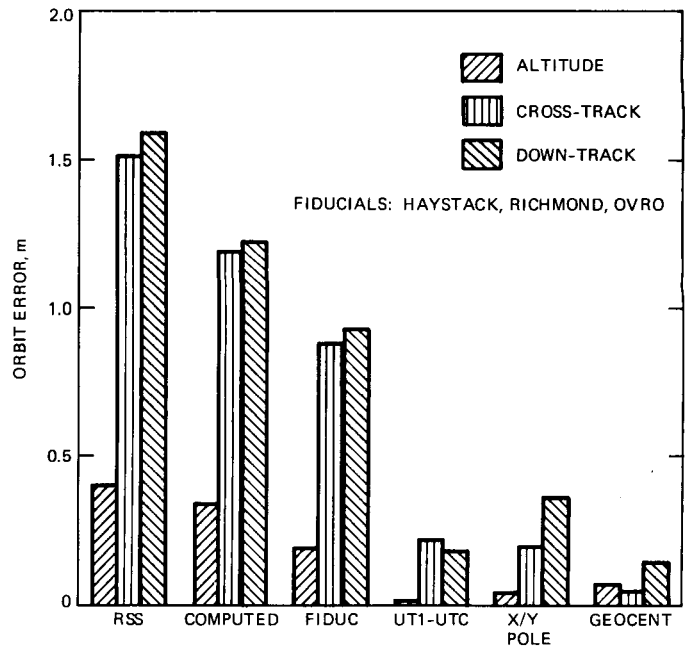


Fig. 14. Consider error analysis for carrier phase (GPS 8) showing relative contributions to the GPS error budget from data noise (computed error), fiducial station coordinate errors, earth orientation uncertainty, and geocenter location error; the configuration of the spring 1985 GPS experiment was used for this analysis

Robust Statistical Methods for Automated Outlier Detection

J. R. Jee
Navigation Systems Section

The computational challenge of automating outlier, or blunder point, detection in radio metric data requires the use of nonstandard statistical methods because the outliers have a deleterious effect upon standard least squares methods. The particular nonstandard methods most applicable to the task are the robust statistical techniques that have undergone intense development since the 1960s. These new methods are by design more resistant to the effects of outliers than standard methods. Because the topic may be unfamiliar, a brief introduction to the philosophy and methods of robust statistics is presented. Then the application of these methods to the automated outlier detection problem is detailed for some specific examples encountered in practice.

I. Introduction: A Specific Problem and a Solution Strategy

Radio metric data must routinely be screened for spurious values which may result, for example, from a temporary malfunction in the instruments gathering the data or from a human blunder in the data collection process. These spurious values typically reveal themselves as outliers, or points which lie outside the normal range of the good data. An efficient way for a data analyst to detect these outliers is to view a computer-generated plot of the radio metric data (or, as in actual practice, a transformed version of them) against their respective time coordinates. The outliers show up in the plot as unusually large deviations from a mean curve followed by the good data. After examining such a plot, the data analyst can delete the outliers from the data set. While this mode of operation for outlier detection has been an acceptable method in the past, it is realized that the expected large increase in the amount of radio metric data to be processed will require excessive amounts of manpower unless steps are taken to automate some portion of the process. The consequent mathematical challenge is that of developing computational algorithms which can perform the job of outlier detection for radio metric data.

Since radio metric data are directly derived from physical quantities such as velocity and range, the mean curve traced by a plot of the good data should be smooth and continuous (in the absence of abrupt dynamic changes); thus, function fitting techniques seem to constitute a natural mathematical tool to use for initial analytical estimation of the mean curve. Then, assuming that the good data fall within a distinct nominal noise band about the mean curve, one can use the analytic function estimate to characterize the outliers as those data which lie outside the band. The special difficulty of this function fitting problem is that the very presence of the out-

liers foils classical function fitting attempts. Indeed, the very reason that outliers need to be removed is that least squares estimation algorithms can be adversely affected by them. More specifically, a function fit by least squares to outlier-contaminated data may not follow the good data well. A common consequence is that the good data deviate as much from the function estimate as the outliers do. To salvage this strategy based on function estimation, it is necessary to employ methods which are robust against the presence of outliers. Fortunately, outlier-robust methods have been a major area of research and development in the statistical community since the 1960s. A major goal of this article is to provide the reader with a brief introduction to these modern statistical methods by way of application to the specific problem described.

A summary of the rest of this article is now given. Section II provides an introduction to some of the philosophy and methods of robust statistics. Sources of information for this section include the seminal article by Huber [1] and his follow-up textbook [2]. Two less theoretical treatments of this topic are given in [3] and [4]. Section III details the application of these methods to the outlier detection problem. The description is devoted more to providing the rationale behind automating outlier detection than to specifying fine algorithmic details. The methods are applied to two sets of radio metric data from the International Cometary Explorer (ICE) spacecraft. Section III also outlines the application of two additional statistical tests which can be found in more standard texts such as [5].

II. Robust Statistical Methods

The operative mode for much of classical statistics is to assume an appropriate probability model and then to employ the optimal procedure for the model. For the purposes of this discussion, the efficiency of a procedure is measured by its theoretical variance. In contrast to classical procedures, robust statistical methods by design rely less heavily upon an appropriate choice of the probability model. Considerations of robustness lead to the development of methods which compromise the requirement of optimal efficiency for the ability to accommodate a range of deviations from a specific assumed model with reasonable, rather than optimal, efficiency. For example, a very common model is the Gaussian density denoted by $\phi(\cdot;\mu,\sigma)$. A class of deviations from this model is given by the mixture densities $(1-\epsilon)\,\phi(\cdot;\mu,\sigma) + \epsilon\phi(\cdot;\mu,N\sigma)$, where N is a large number and ϵ ranges from 0 for no deviation to 1/2 for large deviations. These particular mixture densities with ϵ between 0.01 and 0.1 often provide a more realistic model of real data which tend to be contaminated with outliers. The main fault of classical procedures is that they perform optimally for their intended case $\epsilon = 0$ but often lose efficiency quite drastically as ϵ increases. Robust alternatives remedy this fault by insuring against unacceptably poor performance while not necessarily providing optimal efficiency for any one case.

A. Robust Location and Scale Estimation

The problems of locating the center of a distribution and of determining its scale are the simplest cases for illustration of these concepts. As a specific example, suppose a set of data $\{x_i\}_{i=1}^{n}$ is assumed to come from a Gaussian distribution for which the mean and variance are to be estimated. The classical location estimator is the minimizer of the least squares error criterion $\rho_2(T) = \Sigma(x_i - T)^2$. This estimator is, of course, the sample mean denoted by $\bar{x}_n$. The classical estimate of scale is derived by taking the square root of the sample variance, $\sqrt{s_n^2}$. Suppose that the particular data set has $\bar{x}_n = 0$ and $s_n^2 = 1$, and consider the effect of an additional datum on these estimates. Then the sample mean $\bar{x}_{n+1} = x_{n+1}/(n+1)$ while the sample variance $s_{n+1}^2 = [(n-1)/n] + x_{n+1}^2/(n+1)$. These equations show that a single outlier x_{n+1} can cause the estimates to be arbitrarily large. This is one sense in which the classical estimators of location and scale are not robust against the presence of outliers.

The extreme sensitivity of the sample mean to outliers can be traced to the error criterion from which it is derived. In an effort to balance the total squared error, the sample mean is forced to overcompensate for the one outlier. Consider now changing the form of the error criterion to $\rho_1(T) = \Sigma|x_i - T|$. This criterion is minimized by the median M_n = median (x_i). It is easy to verify that one outlier, or even a small percentage of outliers, has a limited effect upon the location estimate based upon minimizing ρ_1. Thus, the median offers some robustness against outliers and consequently is receiving renewed attention as a location estimator. Similarly, a preferred robust estimate of scale is the Median Absolute Deviation from the sample median, MAD_n = median $\{|x_i - M_n|\}$. For a Gaussian data set, $E[s_n] \approx \sigma$, but $E[\mathrm{MAD}_n] \approx (2/3)\sigma$. Thus, a factor of 3/2 is required to make the MAD_n directly comparable with s_n as an estimator.

B. Robust Linear Regression

In brief review, linear regression analysis is concerned with equations of the form $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $\mathbf{X}$ is a matrix of known quantities, $\mathbf{y}$ is also a known vector often called the vector of observations, β is a vector of unknown parameters to be estimated, and ϵ is a vector of random errors with covariance $E[\epsilon\epsilon^T] \equiv \mathbf{W}^{-1}$. For example, least squares function fitting is one of the problems which may be formulated in the linear regression framework. Classical regression proceeds to a solution by finding the β which minimizes the summed weighted-squared errors $(\mathbf{y} - \mathbf{X}\beta)^T\mathbf{W}(\mathbf{y} - \mathbf{X}\beta)$. The extremal condition

obtained by differential calculus is $\mathbf{X}^T\mathbf{W}(\mathbf{y} - \mathbf{X}\beta) = 0$, from which the solution is quickly obtained. The sensitivity of this least squares solution to outliers is seen by examining the predicted observation vector $\mathbf{y}_{LS} = \mathbf{X}\beta_{LS} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$. This equation shows that if all y-values are held fixed except for y_i, then a change in y_i produces a proportional change in the least squares fitted value $(y_{LS})_i$. Thus, a deviant value can have an arbitrarily large effect upon least squares estimates.

The technique of modifying the error criterion to obtain a robust location estimator is now applied to the regression problem. The least squares criterion may be written as

$$\sum_i (y_i - \mathbf{x}_i\beta)^2 w_i$$

if $\mathbf{x}_i$ denotes a row of $\mathbf{X}$ and $\mathbf{W}$ is a diagonal matrix with nonzero elements w_i. Another standard notation is that for the residual $r_i = y_i - \mathbf{x}_i\beta$. The particular robust error criteria considered here are of the form $\Sigma_i \rho_c(r_i)$, where

$$\rho_c(r) = \begin{cases} c|r| - (1/2)c^2 & \text{if } |r| \geqslant c \\ (1/2)|r|^2 & \text{if } |r| < c \end{cases}$$

where each positive value of c defines a candidate error measure. As c is chosen to be arbitrarily large, the criterion approaches the least squares measure; as c approaches 0, the absolute value criterion is produced. At a minimizer of $\Sigma_i \rho_c(r_i)$, the β must satisfy

$$\sum_{|r_i| \leqslant c} (y_i - \mathbf{x}_i\beta)\mathbf{x}_i^T + \sum_{|r_i| > c} \frac{c}{|r_i|}(y_i - \mathbf{x}_i\beta)\mathbf{x}_i^T = 0$$

This equation reveals that the optimality condition resembles something arising from a weighted least squares problem with two types of weights: weights of 1 for residuals smaller than c and weights of $\sqrt{c/|r_i|}$ for larger residuals. Unfortunately, the proper weighting cannot in general be known before the problem is solved, and the equation as it stands is nonlinear in the unknowns. As for the constant c, setting $c = 1\sigma$ yields an estimator that has good efficiency for independent, identically distributed Gaussian errors while providing robustness against mild contamination. However, the variance of the errors is not always known beforehand, so c may also be another unknown parameter to be determined.

One popular technique of solving the extremal condition equations is called Iteratively Reweighted Least Squares (IRLS). This technique is favored because the algorithm (1) is easy to understand, (2) is easy to implement if a common weighted least squares routine is available, (3) works well in practice, and (4) has fairly well understood convergence properties. For simplicity of exposition, the regression problem is now assumed to have $\mathbf{W}^{-1} = \sigma^2\mathbf{I}$. Then the IRLS algorithm works as follows: (1) the regression problem is solved by the standard least squares method, and residuals from this regression are calculated; (2) the MAD is used to estimate σ, the scale of the residuals; (3) the standard weighted regression equations are solved, where weights of 1 are assigned to residuals less than σ and weights of $\sqrt{\sigma/|r_i|}$ are used otherwise; and (4) steps 2 and 3 are iterated until convergence is achieved. While the theoretical convergence properties of the IRLS algorithm are not completely known, partial results indicate that it may be globally convergent, as is the case in computational practice. Furthermore, it often converges at a linear rate to a minimum of the error criterion.

III. Application to Automated Outlier Detection

The statistically robust estimation methods previously described are now applied to the automated outlier detection problem initially presented. More specifically, the problem is that a set of sequential observations $\{(t_i, y_i)\}_{i=1}^n$ are to be screened for deviant y-values. The observations are quantities derived from radio metric data such as range and doppler. In the particular case to be studied, the derived data are doppler pseudoresiduals generated by the Deep Space Stations while tracking the International Cometary Explorer (ICE) spacecraft. Pseudoresiduals are essentially numerical differences between the observed values and the predicted values.

A. Model Selection

As in any applied mathematical problem, one assumes, either explicitly or tacitly, some model for the physical situation which is to be studied. The two main model assumptions for this application are stated in this section.

Dynamic considerations can be used to derive analytic descriptions of the observations as functions of time. For a single pass of data of several hours' length, polynomials may be used to approximate these functions. From empirical studies, it was found that polynomials of degree $2n$ usually suffice to accurately approximate the mean curve of a set of data of length n hours from a spacecraft in interplanetary cruise. In choosing the degree of the polynomial, it is more desirable to underestimate the required degree than to overestimate it because an overly high order polynomial can fit both good data and outliers. While splines and trigonometric series are other possible candidates for the function approximation, time has not permitted an investigation into their use for this application.

Extensive experience has shown that the nominal noise about the mean curve is well modeled by independent, identically distributed Gaussian random variables. Thus, if good estimates of the variance of the Gaussian noise are obtained, then a 3σ rejection rule should falsely remove only about 1 out of 500 good values. On this basis, an outlier may be defined as a value that is deviant by more than 3σ, where σ is a robust estimate of the scale of the nominal Gaussian noise. From experience, the portion of outliers in a pass is typically between 1 and 10 percent, and they may deviate from the mean curve by as much as a few orders of magnitude above the nominal σ.

B. Algorithm Selection

The mathematical tools required for automated outlier detection have been presented in Section II. In particular, the iteratively reweighted least squares, or IRLS, algorithm is used to fit a polynomial to the outlier-contaminated data according to the error criterion ρ_c. The choice of $c = 1\sigma$ gives an estimator with good statistical efficiency for pure Gaussian errors and resistance to approximately 10 percent outlier contamination. The MAD is used in the IRLS algorithm to estimate the scale. After the IRLS fit is completed, the MAD can be used once more on the residuals from the fit to obtain a final scale estimate upon which the 3σ rejection rule is based.

Figures 1 and 2 illustrate the application of the IRLS algorithm to a specific problem. The data are 2-way doppler pseudoresiduals from the ICE spacecraft. Figure 1 shows a plot of the data along with a standard least squares fourth degree polynomial fit to the entire data set (recall that this is the first step in the IRLS algorithm). Also in the plot are 3σ bands about the least squares fit calculated by the usual standard deviation formula. The two main effects of the outliers on these classical estimators are clearly shown in Fig. 1. First, the extreme outlier near the end of the time segment pulls the least squares polynomial away from the good data. Then the outliers and the oversized residuals near the end of the segment combine to inflate the standard deviation estimate. Figure 2 shows the final IRLS polynomial fit to the data with robust 3σ bands. In this plot, the ordinate has a different scale and the extreme outlier is marked by an arrow at its abscissa. As can be seen, the robust methods give a more agreeable estimate of both the mean curve and the spread of the nominal noise about the curve.

C. Model Verification

If the function estimation and the outlier rejection steps are performed correctly, then according to the assumptions for the model, the remaining residuals should be distributed as Gaussian noise. As is often the case, however, models are not always as accurate as they need to be for the mathematical methods to perform as hoped. Consequently, it is essential that measures for verifying model adequacy are included in this automatic data screening algorithm. The major assumptions which must be checked are that the polynomial does give a good approximation to the function followed by the good data, that the portion of outliers detected is less than 10 percent, and that the data remaining after the editing are Gaussian.

As a specific case, Fig. 3 shows a plot of 1-way doppler pseudoresiduals, an IRLS fifth degree polynomial fit to these data, and a robustly determined 3σ band. Again, arrows at the borders of the plot denote outliers that are out of range. This is an example of model underfitting, as the polynomial does not follow some distinct features of the data. Consequently, the basis for the automated outlier detection program is undermined, and the user of the algorithms should be warned of the unreliability of results obtained in this case.

Underfitting can often be characterized by a tendency for strings of data to lie on one side of the polynomial rather than being more randomly strewn about the fitted curve. Fortunately, there exist standard statistical procedures for verifying the randomness of data based on this idea. A statistical test which considers only the signs of the residuals is called the runs test. The statistic upon which it is based is the total number of runs, denoted by R, of both consecutive positive signs and consecutive negative signs. The exact theoretical mean and variance of R under the hypothesis of randomness are given by

$$E[R] = \frac{2NP}{N+P} + 1$$

and

$$\mathrm{Var}[R] = \frac{2NP(2NP - N - P)}{(N+P)^2(N+P-1)}$$

where N is the number of negative signs and P is the number of positive signs. For data sets of size 50 or more, the random variable R standardized by its mean and variance is approximately a zero mean unit variance Gaussian random variable; thus, the measure of randomness given by R can be calculated easily.

If the retained residuals test negatively for underfitting, then an additional chi-square test for Gaussian behavior can be applied. In a nutshell, this test is based upon measuring the amount of agreement between a histogram of the data and a theoretical histogram determined by a "perfect" Gaussian

sample. If $\hat{\eta}_i, \ldots, \hat{\eta}_k$ denote the number of points in each histogram bin and $\eta_i, \ldots, \eta_k$ the number of points in a "perfect" Gaussian histogram, then the chi-squared test statistic C is given by

$$C = \sum_{i=1}^{k} \frac{(\hat{\eta}_i - \eta_i)^2}{\eta_i}$$

Finally, if the polynomial does not seem to underfit the data and if the data seem to be distributed as Gaussian noise about the polynomial, then a simple count of the percentage of outliers should be performed. Since the algorithms are geared to handle data with 10 percent or less outliers, cases which contain greater than 10 percent detected outliers should be flagged as potential problem sets requiring manual inspection.

IV. Discussion

A prime motivation for employing function fitting in the outlier detection problem for radio metric data is that this seems to mimic the operation of a human data analyst. As stated in the introduction, a data analyst usually bases much of outlier screening upon visual inspection of a plot. The human eye "smooths" the data while ignoring outliers to extract the underlying curve. Function fitting attempts to imitate this process, and robust function fitting in particular is required to accurately produce the underlying curve. After the curve is estimated, the robust 3σ rule provides some analytical basis for outlier detection comparable to an analyst's decision to remove data separated by a "gap" from the bulk of the data spread about the underlying curve.

Another reason for choosing robust statistical methods is that they are fairly easy to understand at the conceptual level. There are more classical statistical procedures for outlier detection in the linear regression framework, but their application in this setting is less straightforward. Typically, classical procedures require more involved statistical reasoning than do robust methods while providing comparable performance; hence the choice for robust methods. For a more detailed comparative discussion, the interested reader can consult [6], which is an extensive survey of methodologies (classical, Bayesian, and robust) for handling outliers in various contexts.

To think that robust statistical methods were first invented in the 1960s is incorrect, since the median and other robust statistics were in use long before then. However, it was not until the 1960s that a unifying framework was established for considerations of robustness. This development prompted considerable theoretical and computational investigation into the subject. Also, while the emphasis in this article has been on robustness against outliers, it should be known that the goals of robust statistics include protection against more than just outliers. A more complete description of this modern statistical methodology can be found in the references. As a note of caution, robust statistical methods are not a new class of foolproof methods which will replace classical methods based on least squares and maximum likelihood. Instead, they constitute another class of methods with its own domain of application alongside those of other statistical methods. One appropriate domain of application is the outlier detection problem of this article. For this case, robust methods should seem not only reasonable but also ideal for the problem.

References

[1] P. J. Huber, "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, vol. 35, pp. 73–101, 1964.

[2] P. J. Huber, *Robust Statistics*, New York: John Wiley, 1981.

[3] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis*, New York: John Wiley, 1983.

[4] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Exploring Data Tables, Trends, and Shapes*, New York: John Wiley, 1985.

[5] I. M. Chakravarti, R. G. Laha, and J. Roy, *Handbook of Methods of Applied Statistics*, New York: John Wiley, 1967.

[6] R. J. Beckman and R. D. Cook, "Outliers," *Technometrics*, vol. 25, pp. 119–163, 1983.

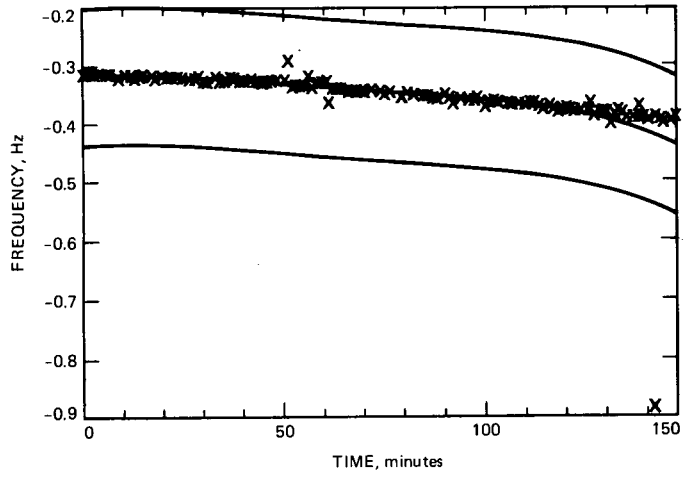


Fig. 1. Fourth degree least squares polynomial and 3 standard deviations based on 2-way doppler pseudoresiduals from ICE. Extreme outlier at time 141 pulls the polynomial away from the rest of data near the end of the time segment and inflates the variance estimate. Time segment begins at 13:18, February 5, 1987.

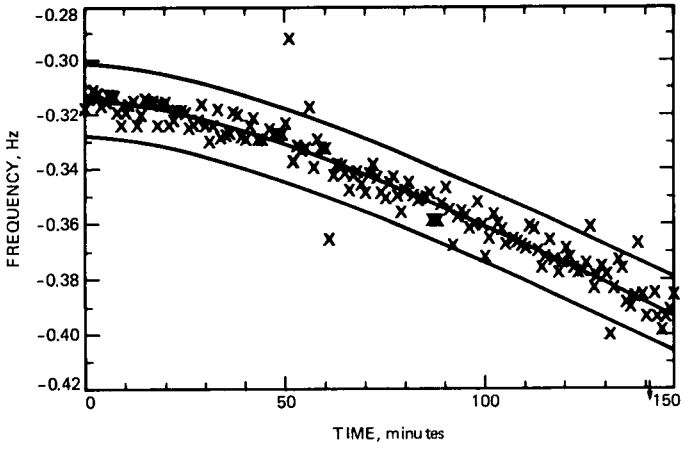


Fig. 2. Fourth degree IRLS polynomial and MAD-based 3σ for 2-way doppler data of Fig. 1. Out-of-range extreme outlier at time 141 is marked by the arrow at the bottom of the graph. Time segment begins at 13:18, February 5, 1987.

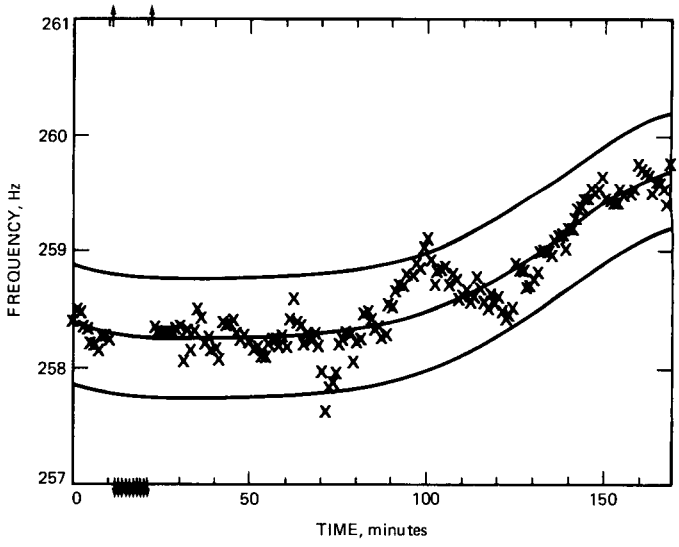


Fig. 3. Fifth degree IRLS polynomial and MAD-based 3σ for 1-way doppler pseudoresiduals from ICE. Out-of-range outliers are marked by arrows. Polynomial underfitting is characterized by the number of runs of deviations of the same sign from the polynomial. Time segment begins at 18:23, February 7, 1987.

Precise Near-Earth Navigation With GPS: A Survey of Techniques

T. P. Yunck, S. C. Wu, and J. Wu
Tracking Systems and Applications Section

The tracking accuracy of low earth orbiters (below ~3000 km altitude) can be brought below 10 cm with a variety of differential techniques that exploit the Global Positioning System (GPS). All of these techniques require a precisely known global network of GPS ground receivers and a receiver aboard the user satellite, and all simultaneously estimate the user and GPS satellite orbits. Three basic approaches are the geometric, dynamic, and non-dynamic strategies. The last combines dynamic GPS solutions with a geometric user solution. Two powerful extensions of the non-dynamic strategy show considerable promise. The first uses an optimized synthesis of dynamics and geometry in the user solution, while the second uses a novel gravity-adjustment method to exploit data from repeat ground tracks. These techniques will offer sub-decimeter accuracy for dynamically unpredictable satellites down to the lowest possible altitudes.

I. Introduction

Tracking requirements for earth sensing satellites, particularly altimetric satellites, are becoming increasingly stringent, reaching the decimeter level for several missions proposed for the 1990s. NASA's Ocean Topography Experiment (Topex) [1]–[3], scheduled for launch at the end of 1991, has a goal of 13 cm altitude accuracy but would benefit from an accuracy level comparable to the 2.5 cm precision of its radar altimeter. A number of similar missions, including the Navy Remote Ocean Sensing System (NROSS) [4], the European Space Agency's Earth Remote Sensing-1 (ERS-1) [5], [6], and a series of altimetry experiments planned for NASA's Earth Observing System (EOS) [7], [8], are also seeking decimeter altitude accuracy.

Topex will carry an experimental tracking system which exploits the U.S. Defense Department's Global Positioning System (GPS) [9], [10]. The basic technique, sometimes called differential GPS, makes use of a high performance GPS receiver on board the orbiter and a small network of precisely located receivers on the ground, distributed roughly evenly around the globe. All receivers continuously observe the visible GPS satellites, making measurements of accumulated RF phase and one-way range at roughly 1.2 and 1.6 GHz (Fig. 1). The one-way range is also called "pseudorange" because it consists of true range plus the time offset between the transmitter and receiver clocks. Orbiter and ground measurements are then combined and processed to recover the orbiter and GPS satellite states in the reference frame defined by the ground network [11]–[19].

For the Topex demonstration, the reference network will include DSN tracking stations in California, Spain, and Australia and at least three complementary sites operated by the Defense Mapping Agency (DMA). A map of the hypothetical

sites used in the error analysis presented here, along with a typical one-orbit Topex ground track, is given in Fig. 2. At Topex launch, ground site relative positions are expected to be known to about 3 cm and their geocentric positions to about 5 cm.

Our purpose here is to evaluate different strategies for applying differential GPS carrier and pseudorange data to determine the user orbit. One requirement of all GPS-based strategies intended to achieve decimeter accuracy is that a joint solution be performed for the user and GPS satellite orbits. If GPS orbits are left unadjusted, user position accuracy is generally limited at the meter level by the a priori GPS orbit error. Because there will be at least 18 GPS orbits to estimate along with the user orbit, considerable data strength is needed for a high accuracy solution.

II. Three Basic Strategies

We begin with three fundamental GPS-based differential tracking strategies: a purely geometric strategy, a fully dynamic strategy, and a combined strategy in which GPS orbit solutions are dynamic and the user solution is geometric. Before presenting the details, a discussion of data combining will be helpful. All of these strategies can use what are called undifferenced, singly differenced, or doubly differenced GPS observables. These correspond, respectively, to the cases in which (1) all clock behavior is modeled over time; (2) GPS clocks are eliminated and only receiver clock behavior is modeled over time; and (3) all clocks are eliminated (or solved for) at each time step. In general, the less differencing applied, the greater the data strength available for orbit solutions. When receiver oscillators are unstable, however, they can seriously degrade the solution by introducing systematic errors. Where possible, it is advantageous in such cases to eliminate oscillator effects through differencing. The analysis presented here assumes the use of doubly differenced data in order to remove oscillators as a possible source of error. Results will invariably improve if oscillators are sufficiently stable to permit less differencing.

A. The Geometric Strategy

This is the differential analog of classical GPS-based point positioning in which a user makes pseudorange measurements to four or more GPS satellites, obtaining a quick geometric solution for position and time offset from GPS time. The conventional (non-differential) user is dependent upon a priori knowledge of GPS satellite positions and time offsets, which for most users are expected to be in error by roughly 5 m. This limits final position error to 10–15 m. In the simple differential approach (Fig. 1), the user and a reference receiver make pseudorange measurements to a common set of at least four satellites, permitting a geometric solution for the baseline and the time offset between receivers. GPS orbits are left unadjusted, but cancellation of common GPS clock and orbit errors improves user position accuracy to one or two meters.

In the general approach, a network of reference receivers and the user receiver observe the GPS satellites. With sufficient measurements, the user position and all GPS positions can be determined geometrically with respect to the reference ground network. To illustrate, consider a set of four GPS satellites and assume the use of doubly differenced measurements. Including the user, there are 5 satellites or 15 position components to estimate, requiring at least 15 independent doubly differenced measurements. Since each user-to-ground baseline yields three double differences, five baselines and thus five reference receivers are needed, all viewing the same four satellites. If these conditions are met, the user and GPS orbits can be simultaneously estimated, and performance will be limited primarily by measurement precision and observing geometry.

With the current generation of receiving equipment, pseudorange measurements are typically precise to 0.5–1.0 m over one second. This limits instantaneous position accuracy to a meter or worse, depending upon observing geometry. In any geometric strategy, the pseudorange precision limitation can be overcome by introducing GPS carrier phase, continuously counted, and subtracting it from pseudorange, thereby removing receiver and GPS dynamics and permitting pseudorange to be averaged over time. This procedure is known as "smoothing pseudorange against the carrier." More generally, position change obtained from carrier phase measurements can be subtracted from successive position solutions obtained with pseudorange, permitting the averaging of position solutions over arbitrarily long time periods despite frequent switching of GPS satellites. By this means, the general geometric strategy can in principle be made to deliver decimeter accuracy for any low orbiter. In reality, however, an impractically large receiver network is needed to maintain strong determination of all parameters, and therefore the purely geometric approach cannot compete with more efficient alternatives.

B. The Dynamic Strategy

The most familiar of the alternatives is the classical dynamic formulation in which the state parameters at a single epoch are estimated for all satellites using an extended arc of data [11]–[13]. Observations at different times are related to the epoch state parameters by integrating the equations of motion, a process requiring accurate models of the observing system and of the forces acting on the satellites. Errors in the dynamic models naturally introduce errors in the epoch state solution.

In general, the further in time an observation is from the solution time, the greater the expected error from dynamic mismodeling. Consequently, the effect of force model errors tends to increase with increasing arc length.

Compared with the general geometric strategy, this approach vastly reduces the number of estimated parameters, thus increasing data strength and permitting far fewer reference receivers. Moreover, by introducing dynamic constraints, this technique permits solutions that are impossible geometrically, such as the determination of satellite positions from carrier phase (position change) measurements alone. The dynamic technique can therefore be used with carrier phase, pseudorange, or both. These considerable advantages are gained in return for a dependence upon dynamic models, with a consequent vulnerability to modeling errors.

Two recent Topex studies illustrate the importance of accurate models with the dynamic technique. Figure 3 is taken from a covariance study of the altitude error for a single Topex orbit using carrier phase data. Key assumptions include: Topex viewing 4 GPS satellites; integrated doppler data with 0.4 cm noise and 5 min data interval; 6 ground sites known to 5 cm in each component; zenith troposphere calibration accurate to 1 cm; 18-GPS constellation with 4 m *a priori* ephemeris error. The most critical assumption is the gravity error model, which consisted of the differences between 21 selected coefficients taken from two gravity models, GEM10 and GEM10B [20], [21]. These differences were further reduced by 50 percent to account for expected model improvements before Topex launch. Figure 4 presents a similar analysis (although for a different orbit–note the changed data noise error) which used a gravity model of more than 400 terms differenced between GEML2 [22] and GEM10, this time without the 50 percent reduction. This reflects the approximate accuracy of the best gravity models in the early 1980s. We can see that for a fully dynamic solution strategy, a significant model improvement is needed to reach the Topex 13-cm goal. Efforts over the last several years at the Goddard Space Flight Center and at the University of Texas have led to considerable progress on the gravity model and are expected to result in the achievement of the required accuracy by the time of Topex launch [23].

Although the dynamic strategy looks promising for Topex, were it to be applied at much lower altitudes, such as the 250–350 km typical of shuttle flights, errors would soar. Gravity error would grow substantially and yet would be far surpassed by the greatly increased error in modeling atmospheric drag. In the case of the shuttle, additional complications would arise from the irregular effects of maneuvering and venting. Final orbit error could reach hundreds of meters. For high accuracy at very low altitudes, one is therefore led back to a geometric approach.

C. The Non-Dynamic Strategy

This technique was proposed in 1985 [17] to address specific weaknesses of the two previous techniques. To eliminate significant modeling errors, the user position is once more determined geometrically, with a new and independent solution derived at each time point. To prevent the proliferation of estimated parameters, as occurs with the fully geometric approach, the GPS satellite states are obtained at a single epoch by a conventional dynamic approach. This strategy can be thought of as a classical epoch state dynamic formulation with the user state modeled as process noise with zero correlation time.

Although dynamics appear in this approach, they do so only for the high altitude GPS satellites, for which dynamic modeling errors are negligible. Dynamic treatment of GPS orbits sharply reduces the number of estimated parameters, permitting use of a small reference network. Since the user solution is geometric, the customary dynamic error sources–gravity, drag, solar radiation, maneuvers, and venting, to name a few–are eliminated. In recognition of the geometric user solution, we refer to this technique as *non-dynamic* tracking. Note that in order to obtain a non-dynamic position solution, the pseudorange data type is required. If pseudorange alone is used, however, performance will again be limited at the meter level by the relatively high pseudorange measurement error. The real power of this technique emerges when continuous carrier phase is introduced, allowing the smoothing of position solutions against observed position change. Several hours of smoothing can reduce the contribution of data noise to position error to a few centimeters.

An example from the extensive Topex error studies [17], [18] is shown in Fig. 5. In this case, Topex is assumed to view up to six GPS satellites at once, rather than four. The data noise on the dual frequency, doubly differenced pseudorange is assumed to be 10 cm after a 5 min integration. Other relevant assumptions remain the same. Over a 4-hr data arc, the average altitude error is 7.3 cm, with two peaks of about 12 cm. Several features distinguish this result fundamentally from the dynamic results. Because dynamic model errors are absent, this accuracy can be maintained down to the lowest satellite altitudes (roughly 160 km) without concern for unmodeled forces or possible maneuvering of the vehicle–so long as contact with GPS is not disrupted. Moreover, since there are no user dynamic models to compute, the solution procedure is considerably simpler and faster than with a dynamic approach.

The pseudorange data noise assumed in Fig. 5 corresponds to a single-channel precision of approximately 40 cm in one second. The new "Rogue" receiver now undergoing field tests at JPL improves on this by about a factor of two [15], [24],

while some current commercial receivers have about double this error. Doubling the data noise in Fig. 5 increases the average altitude error to 8.5 cm; halving it reduces the error to 7.0 cm. The receiver to be carried aboard Topex, now being developed by Motorola, is expected to have a pseudorange measurement noise somewhat better than the 40 cm assumed here. It should be noted that for these results to be strictly valid, sources of systematic error such as multipath and instrument delay variations must be contained so that after four hours of averaging their effect is below that of the data noise.

A final point is worth noting. This study assumed a six-receiver reference network and a flight receiver observing up to six satellites. Ground receivers were assumed to track all satellites above 10 degrees. If the flight receiver is restricted to viewing four satellites, as may be the case for Topex, the observing geometry frequently degrades sharply or even breaks down altogether, and errors soar. Enlarging the ground network to 15 sites does not fully restore performance. A more robust strategy to deal with weak geometry is presented in Section IV.

III. Carrier Range

There is a data type called "carrier range" which is sometimes recoverable from differential GPS observations. Carrier range is obtained by determining the exact number of full cycles in the carrier phase observable differenced between two receivers. It is, in effect, a differenced pseudorange having the sub-centimeter precision of carrier phase measurements. The process of determining the integer cycle count is called cycle ambiguity resolution or "bias fixing," and a variety of techniques have been devised to carry it out [15], [25], [26]. Bias fixing is a demanding task that currently can be reliably achieved only between fixed ground sites no more than a few hundred kilometers apart. Several groups are now trying to achieve bias fixing over continental distances.

Carrier range performance can often be approached with the combined carrier phase and pseudorange data types. Carrier phase has the precision of carrier range without the position information; pseudorange has the position information without the precision. Over long data arcs, pseudorange error can be averaged down to bring the effective data noise near that of carrier range. Other errors tend to mask the remaining difference.

Consider, for example, the non-dynamic analysis with combined data types shown in Fig. 5. The same case using carrier range, shown in Fig. 6, yields an average improvement of 1.3 cm. (To optimize performance in Fig. 6, three ground sites were adjusted. Without this refinement, there is no net improvement.) When the assumed pseudorange error is halved to correspond to that of the best GPS receivers, the carrier range advantage is reduced to 1.0 cm. Thus, if bias fixing proves to be unattainable over long distances, the combined data type may offer a practical alternative. Again we must stress that in precise applications of pseudorange, multipath must be carefully controlled.

IV. Two Advanced Strategies

The non-dynamic strategy has an important limitation: Performance is strongly dependent upon the momentary observing geometry, as evidenced by the error fluctuations in Fig. 5. The momentary observing geometry depends in turn upon the number and arrangement of ground receivers, the receiver viewing capacities, and the GPS satellite constellation. Loss of a key ground site, GPS satellite, or user channel can cause the solution to degrade sharply or even fail altogether. One way to address this is by reintroducing user dynamics, appropriately weighted according to model quality, while preserving the essential geometric technique. With this approach, which is developed in detail below, user dynamics smooth the solution through geometric trouble spots while adding information and strength throughout.

A. The Reduced Dynamic Strategy

To put this idea into practice, we return to the dynamic formulation and introduce a non-dynamic component to the solution by incorporating added process noise in the user force models. The GPS state solution remains fully dynamic, while the user solution is, in general, partly dynamic and partly non-dynamic. Widely different solution characteristics, ranging from fully dynamic to non-dynamic, can be achieved simply by varying the parameters defining the process noise. The solution can be optimized for a particular combination of geometric strength and dynamic model accuracy by careful tuning of those parameters.

We present this "reduced-dynamic" technique mathematically in a Kalman sequential filter formulation. This involves two steps: a *time update*, which makes use of a state transition model to propagate the satellite state estimate and covariance from one time batch to the next, and a *measurement update*, which incorporates a new batch of measurements. These two steps alternate until all data batches are incorporated.

In the *time update*, let $\hat{\mathbf{x}}_j$ be the user satellite state estimate at time t_j using data up to the time t_j, and $\tilde{\mathbf{x}}_{j+1}$ the predicted state estimate at time t_{j+1} using data only up to t_j; let $\Phi_x(j+1, j)$ denote the state transition from t_j to t_{j+1}. We now introduce process noise parameters $\mathbf{p}$ representing a fictitious 3-D force on the user satellite. This gives us the following dynamic (state

transition) model for the augmented state $\mathbf{X} = [\mathbf{x}, \mathbf{p}]^T$ and its associated covariance P [27]:

$$\tilde{\mathbf{X}}_{j+1} = \Phi_j \hat{\mathbf{X}}_j + B\mathbf{w}_j \tag{1}$$

and

$$\tilde{P}_{j+1} = \Phi_j \hat{P}_j \Phi_j^T + BQ_jB^T \tag{2}$$

where

$$\Phi_j = \begin{bmatrix} \Phi_x(j+1, j) & \Phi_{xp}(j+1, j) \\ 0 & M_j \end{bmatrix} \tag{3}$$

and

$$B = \begin{bmatrix} 0 \\ I_p \end{bmatrix} \tag{4}$$

In Eq. (3), $\Phi_{xp}(j+1, j)$ is the transition matrix relating $\tilde{\mathbf{x}}_{j+1}$ to the process noise parameters $\mathbf{p}_j$, and M_j is a 3×3 diagonal matrix with its ith element

$$m_i = \exp\left[-(t_{j+1} - t_j)/\tau_i\right] \tag{5}$$

In the above equations, $\mathbf{w}_j$ is a white noise process of covariance Q_j which is also diagonal with its ith element $q_i = (1 - m_i^2)\,\sigma_i^2$, and I_p is a unit matrix. Both the steady-state uncertainty σ_i and the correlation time constant τ_i can be selected to be the same for all i in this application; thus we will drop the subscript i. The relative weighting of the dynamics is varied by selecting different values for the a priori uncertainty σ_0, the steady-state uncertainty σ, and the correlation time τ for these process-noise parameters. Increasing τ and decreasing σ_0 and σ increases the weight on the dynamic information. When $\tau \to \infty$, $\sigma \to 0$, and $\sigma_0 \to 0$, the technique reduces to conventional dynamic tracking; when $\tau \to 0$, $\sigma \to \infty$, and $\sigma_0 \to \infty$, it becomes non-dynamic tracking. It follows that an optimally tuned reduced-dynamic solution must always be as good as or better than both the fully dynamic and the non-dynamic solutions.

The model for a *measurement update* in the reduced-dynamic technique is the same as that for conventional dynamic or non-dynamic tracking, with the exception that $\mathbf{X}$ and P are now associated with the augmented state. Thus,

$$\hat{\mathbf{X}}_j = \tilde{\mathbf{X}}_j + G_j(\mathbf{z}_j - A_j \tilde{\mathbf{X}}_j) \tag{6}$$

and

$$\hat{P}_j = \tilde{P}_j - G_j A_j \tilde{P}_j \tag{7}$$

where $\mathbf{z}_j$ is the measurement vector at time t_j; A_j is the matrix of the corresponding measurement partials with $\mathbf{X}_j$; and G_j is the Kalman gain given by

$$G_j = \tilde{P}_j A_j^T (A_j \tilde{P}_j A_j^T + R_j)^{-1} \tag{8}$$

with R_j being the error covariance of $\mathbf{z}_j$. These models are formulated in terms of current state for clarity. A pseudoepoch state U-D factorized formulation [27] has been implemented in the GPS error analysis software known as OASIS (Orbit Analysis and Simulation Software) developed at JPL [28].

1. Performance analysis. Further insight into the reduced-dynamic technique can be gained by comparing its performance in all its key forms, including fully dynamic, optimized reduced dynamic, and non-dynamic. For this analysis we assume a constellation of 18 GPS satellites in six orbit planes, with pseudorange and accumulated carrier phase measurements acquired by Topex and six ground receivers every 5 minutes. Data noise, after dual frequency combination, is set at 5 cm and 0.5 cm, respectively, for the two data types. Carrier phase biases, which must be estimated in the solution process, are given large a priori uncertainty. A 2-hour data arc covering a full Topex orbit is used in all cases. The Topex ground track and positions of the six ground sites are again those of Fig. 2. Other error sources considered are given in Table 1.

To eliminate systematic oscillator error, the clocks at all GPS satellites and at all but one ground site are modeled as white process noise and estimated at each time step. This is a general form of the double differencing technique discussed in Section II. Earth gravity error is represented by a lumped model of a 20×20 field derived by taking 50 percent of the difference between two earth models, GEM10 and GEML2. Comparisons of this representation with the estimated accuracies of the best current gravity models suggest that this error model somewhat overstates the true error by perhaps 20 or 30 percent. The 1-cm zenith troposphere error given in Table 1 assumes the use of a water vapor radiometer at each ground site. Both Topex and the ground receivers are assumed to observe all GPS satellites within their fields of view (typically 6 or 7) unless otherwise stated. The relative performance of the different filter methods is assessed by comparing the estimated Topex altitude errors over the entire 2-hour span. For this, the state covariances of Topex are first smoothed backward and are then mapped to all time points when data

are taken. Comparison is made between the root-mean-square (RMS) errors calculated over these time points.

In preliminary studies not shown here, the properties of the two limiting cases of the reduced-dynamic technique were confirmed. With τ set to 0 and both σ_0 and σ set to a large number, the error estimate, as expected, approaches the solution derived from a separate non-dynamic formulation. When τ is large and both σ_0 and σ are set to 0, the purely dynamic result is produced. Now let's examine a series of intermediate values for τ, σ_0, and σ. The results, in general, will vary with the batch-to-batch uncertainty $\sigma_{bb} \equiv (1 - m^2)^{1/2}\sigma$ rather than with the steady-state uncertainty σ and τ individually. Therefore, in the following analysis a constant τ = 15 minutes is used; only $\sigma_0 = \sigma$ is varied to yield a nearly optimal solution.

Figure 7 shows the Topex altitude error as a function of the percentage of the GEM10–GEML2 error for various values of σ. This includes the results for dynamic tracking ($\sigma = 0$, $\tau \to \infty$) and non-dynamic tracking ($\sigma \to \infty$, $\tau = 0$). It is clear that for any finite dynamic model error (in this case dominated by the gravity), a range of σ exists with which Topex altitude error is lower than with either the dynamic or the non-dynamic solution. In other words, the reduced-dynamic technique is superior provided that the dynamic model is properly weighted.

In Fig. 8, the reduced-dynamic solution is compared with the dynamic and non-dynamic solutions for three different observing capacities for the GPS receiver on board Topex: simultaneously observing 4 GPSs, 5 GPSs, and all GPSs (typically 6, and seldom more than 7) above the Topex horizon, which is defined to be 90 degrees from zenith. In the cases with restricted receiver capacity, satellites are selected to minimize switches over the observing period, thereby maximizing continuity in carrier phase measurements while maintaining good observing geometry (low PDOP). The gravity error is fixed at 50 percent of the difference between GEM10 and GEML2, which is roughly the level of our current uncertainty. A near-optimum weight (σ = 0.5 μm/sec^2) is used for the reduced-dynamic solution in all three cases.

With a Topex receiver observing all visible satellites, the geometry is always good and non-dynamic tracking is powerful; incorporating the extra dynamic information using the reduced-dynamic technique improves the accuracy by only 1 cm. An improved gravity error, perhaps achieved through gravity tuning with Topex data, would of course improve the reduced-dynamic technique. At the other extreme, with Topex observing only 4 satellites, the geometry is often poor, and non-dynamic tracking performance is much worse than that of dynamic tracking. The reduced-dynamic combination therefore offers little improvement over purely dynamic tracking. If the gravity error is doubled, however, as in the case of a lower orbit, dynamic tracking error nearly doubles, to 24 cm, while reduced-dynamic performance degrades only moderately, to 16 cm. The third case, Topex observing 5 satellites, falls between these extremes. Dynamic tracking and non-dynamic tracking give 12 and 16 cm, respectively. The reduced-dynamic combination improves this to 9 cm, illustrating the clear advantage of reduced-dynamic tracking when dynamic and non-dynamic performance levels are comparable.

Dynamic tracking naturally yields higher error over regions where gravity is poorly known, e.g., over ocean basins. Non-dynamic tracking, on the other hand, is vulnerable to poor observing geometry. In the reduced-dynamic combination, the two techniques complement one another; the weakness of each is covered by the other's strength, and the solution is better balanced. This is illustrated in Fig. 9, which compares Topex altitude determination accuracy over the whole orbit (2 hours) using the three techniques. A Topex receiver observing up to 5 satellites and the 50 percent GEM10–GEML2 gravity error are assumed. Both dynamic and non-dynamic solutions show peak errors of 25 cm or higher at some points. The reduced-dynamic solution using a near-optimum weight (σ = 0.5 μm/sec^2) smooths these peaks and remains below 13 cm for the whole 2-hour period. Reduction of the error peaks is the result of a near-optimum trade of state transition information between the dynamic and non-dynamic approaches,

At times when the transition information of one approach is poor, the least-squares estimation filter shifts weight to the other, minimizing the overall error. To illustrate this, Fig. 10 breaks down the Topex altitude error at three times when either the dynamic or the non-dynamic technique performs poorly. The near-optimum trade of state transition information in the reduced-dynamic solution has yielded a more uniform contribution from all error components.

The robustness of the reduced-dynamic technique was further demonstrated in an independent study by Williams [8] that investigated an example of temporary data outage during which the non-dynamic technique failed to produce a useful solution. In the same situation, the reduced-dynamic technique automatically shifts full weight to the dynamics, with the result that there is no noticeable loss of accuracy during the outage provided it does not last too long.

Up to this point, a data arc of only 2 hours has been used. In general, with any solution technique, the effects of data noise, station location, and troposphere are naturally reduced as the data arc length increases. With conventional dynamic tracking, the effects of dynamic error typically increase with arc length and eventually begin to dominate. As a result, it is

usually necessary to choose a compromise arc length that balances data errors and dynamic errors. With the reduced-dynamic technique, however, this is not the case. As the arc length increases, the data strength increases; as a natural consequence of the estimation process, the reduced-dynamic technique then shifts greater weight to geometry in such a way as to maintain the balance between data errors and dynamic errors. In effect, dynamics are continuously deweighted as the arc length increases in order to take advantage of the growing data strength. No change in σ is needed for this, since the optimum σ applies to a specific dynamic model error independent of data span. As a result, with optimal weighting the overall performance will not degrade, and will generally improve, with increased data span.

To demonstrate this, Fig. 11 compares the Topex altitude error using 2-hr and 4-hr data spans. The longer data span reduces the error to less than 10 cm over the entire 2-hr period. The RMS altitude error is 7 cm, as compared to 8.9 cm for the 2-hr tracking. An examination of the error breakdown shows a reduction in gravity error as well as in other errors. Although data spans longer than 4 hours have not been studied, it is expected that the error will reduce monotonically with data span. Owing to reduced weight on the dynamic model with longer data span, a reduced-dynamic solution will gradually become a non-dynamic solution as the span is increased. Note that this is true only under the assumption that a fixed, predetermined dynamic model is used, independent of the data arc length. If the gravity model is steadily improved through tuning or other efforts, the optimum weight for a given data span will shift toward a more strongly dynamic solution.

2. Weighting the dynamic model. To benefit fully from the reduced-dynamic technique, the weight on the dynamic model, specified by σ with any adopted τ, must be near its optimum value. However, the sensitivity to a departure from optimum weighting appears to be low. This is illustrated in Fig. 12, which plots the orbit error as a function of the level of gravity error for five different values of weighting, assuming a 5-satellite Topex viewing capacity (compare Fig. 7). The curves for $\sigma = 0.125\ \mu m/sec^2$ and $\sigma = 2\ \mu m/sec^2$ intersect at a gravity error of about 42 percent of GEM10–GEML2. The optimum weight at that point is $\sigma = 0.5\ \mu m/sec^2$. These three curves form a shallow triangle which lies nearly horizontal. Even with σ a factor of 4 from the optimum, Topex altitude error increases by only 0.6 cm. In other words, performance is insensitive to σ near the optimum.

For a particular application, the proper weight can be estimated in advance through a covariance analysis using a realistic dynamic error model. Misjudgment of dynamic error will, of course, yield a suboptimal weight. Care can be taken, however, to minimize the resulting error. The following is a simple strategy: Use a nominal dynamic error model to predict the performance of both dynamic and non-dynamic techniques. If either technique is clearly superior to the other—say, by a factor of 3 or more—the slight improvement that would result from combining the two approaches with the reduced-dynamic technique may not justify the extra effort (although the enormous advantage of having dynamics to fill in when geometry fails, say, as a result of a hardware failure, argues against eliminating dynamics altogether). If neither technique is clearly superior, a weight departing from the predicted optimum in a direction favoring the non-dynamic technique (i.e., with a larger σ) should be chosen. The "biased" weighting reduces the more damaging effect when gravity error actually is larger than expected.

This is illustrated in Fig. 13, which shows Topex altitude error, again for three reduced dynamic weights, over a wider range of gravity error. The dotted line shows the possible performance with optimal weight. Suppose the nominal gravity error is 42 percent of the difference between GEM10 and GEML2. The weight $\sigma = 0.5\ \mu m/sec^2$ is nearly optimal at this value. If the actual gravity error is, for example, a factor of 2.4 larger (100 percent GEM10–GEML2), this weight would degrade the altitude determination from 9.7 cm at its true optimum ($\sigma = 2\ \mu m/sec^2$) to 12.3 cm. A weight favoring the dynamic approach, $\sigma = 0.125\ \mu m/sec^2$, would raise the error to 19.1 cm. If instead the actual gravity is a factor of 2.4 smaller (18 percent GEM10–GEML2), the weight favoring the non-dynamic approach would degrade the solution from the possible 6.2 cm to 9.0 cm. The nominal optimum weight of $\sigma = 0.5\ \mu m/sec^2$ would yield 7.7 cm, which is only marginally better. Therefore, a biased weight favoring the non-dynamic approach (using a larger σ) is preferable when the level of dynamic error cannot be well determined.

B. A Gravity Adjustment Strategy

We now turn to a somewhat different approach in the form of a strategy specialized for a particular class of orbit and for strictly non-real-time application. The orbit must feature a regularly repeating ground track and must be at an altitude between roughly 600 and 3000 km, where gravity is the dominant dynamic error. Orbits for most earth observation satellites requiring high accuracy tracking, including Seasat, Topex, and ERS-1, fit these conditions. The technique features a novel gravity-adjustment strategy which exploits the special character of the orbit to achieve accuracy and efficiency.

Ordinarily the earth's gravity field is represented by a spherical harmonic expansion, and ordinarily several hundred coefficients are needed to support precise dynamic tracking of a low orbiter. Gravity "tuning," which is frequently carried

out to improve orbit accuracy, involves the adjustment of selected harmonic coefficients as part of the orbit determination process. Resonant components with large effects on a particular satellite orbit can thereby be improved, giving the desired orbit improvement.

Here we dispense with the harmonic representation and substitute a set of local parameters, or "bins," spaced evenly around each full ground track. For the fitting process, data from a large number of repeat ground tracks are collected in an ensemble. The user states at the beginning of each arc are estimated together with a single set of 3-D position corrections. These corrections correspond to orbit perturbations in each bin, common to all repeat arcs, that are due to gravity mismodeling, as illustrated in Fig. 14. Because the gravity perturbations felt by the user are the same for repeat ground tracks, collecting repeat orbits in an ensemble permits accurate recovery of local gravity effects by averaging random and other non-repeating errors. The number of parameters needed for the entire globe, and hence the effective gravity resolution, is roughly the same as with a harmonic expansion. With this approach, however, only the relatively few parameters pertaining to a particular ground track are dealt with at one time.

The mathematical details are given in [29] and [30]. Here we present a brief summary. Let $r(t)$ be the deviation of the orbiter position from a nominal trajectory and let t_j^i represent the j^{th} time point of the i^{th} trajectory in the ensemble, where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, M$. Then we can write the linearized expression

$$r(t_j^i) = \frac{\partial r}{\partial r_0} r_0^i + \frac{\partial r}{\partial v_0} v_0^i + \frac{\partial r}{\partial p^i} p^i + d_j \qquad (9)$$

where r_0 and v_0 are the deviations in epoch position and velocity, p^i represents the effects of all non-gravitational dynamic parameters, and d_j is the local parameter, common to all arcs, representing the position correction due to gravity mismodeling at that point. Solving for the d_j in the fitting process constitutes the gravity adjustment performed by this technique. We can write the above equation in matrix form as

$$\mathbf{R}^i = (V\mathbf{x}_0)^i + \mathbf{d} \qquad (10)$$

where $\mathbf{R}^i$ is the vector of position corrections for each time point in the i^{th} arc, $\mathbf{x}_0$ is the correction to the epoch state vector for the i^{th} arc, V is the corresponding matrix of variational partials, $\mathbf{d}$ is the arc-independent vector of position corrections due to gravity mismodeling, and the non-gravitational terms $\mathbf{p}$ have been omitted. The a priori covariance of $\mathbf{d}$ can be derived from the gravity field used for the nominal trajectory by means of a transformation matrix of variational partial derivatives, as described in [30]. Combining measurement data from the multiple data arcs, we can write the standard regression equation

$$\mathbf{z} = A\mathbf{x} + \mathbf{n} \qquad (11)$$

where $\mathbf{z}$ is the measurement vector, $\mathbf{n}$ is the data noise vector, $\mathbf{x}$ is the vector of parameters (including GPS states) to be estimated, and A is the matrix of measurement partials. Note that both A and $\mathbf{x}$ can be partitioned into arc-dependent and arc-independent parts. An efficient method of solving the partitioned regression equation by application of the Householder transformation is given in [29].

Although perhaps it isn't immediately evident, this is an extension of the non-dynamic strategy to repeating data arcs. As presented, the technique yields solutions for satellite epoch states in each arc plus arc-independent user position corrections in each gravity bin. In the degenerate case of a single data arc, however, the notion of arc-independent parameters collapses, and we can obtain a simple set of geometric position corrections—the $r(t_i)$ defined above. This, in essence, is the non-dynamic technique.

Results from a covariance analysis of this technique applied to the Topex example are presented in Figs. 15 and 16. The analysis examines ensembles of repeat tracks ranging in number from one to 100 and employs the same assumptions used in the analyses of Section 4, Subsection B. Figure 15 presents the total error from all sources, while Fig. 16 gives the error from data noise alone. As expected, the single-track case, represented by the top curve in each figure, gives the same result as the non-dynamic technique (not shown) for that arc, an average altitude error of about 8 cm. Performance improves rapidly with the first few additional arcs and then levels off at about 5 cm for a large number of arcs. The limiting error is the 5 cm error in each component assumed for the ground receivers, which contributes 4.9 cm of the total. Note in Fig. 16 that the data noise contribution continues to improve with added data, falling below 1 cm for 100 arcs. Current estimates for the error of our best known global reference points are in the range of 10 cm per component. Projections are that by the early to mid-1990s, many global points will be determined to 2 or 3 cm with improved VLBI and satellite laser ranging techniques. This would lead to a corresponding improvement in the performance shown in Fig. 15.

A further application of the repeat track concept lies in the improvement of gravity models. Finite time-differencing of the arc-independent position corrections yields local gravity

parameters. These can be collected from all ground tracks in the repeat sequence, and by application of suitable transformations, a conventional harmonic gravity model can be produced which is tailored for the particular orbit. Because data from the repeat ground tracks have been reduced to a small set of parameters, the final transformation to a global representation is computationally efficient. Resolution and accuracy will be essentially the same as with other techniques using the same set of data. Analysis for the Topex orbit indicates that with 100 repeat arcs for each ground track, a gravity field at Topex altitude can be recovered with an accuracy of 0.04 milligal and a resolution equivalent to a 12 X 12 spherical harmonic field. The accuracy will improve for lower values of degree and order and will fall off rapidly for higher values. For satellites at lower altitudes, much greater sensitivities can be achieved, particularly for terms of higher degree and order.

V. Summary and Conclusions

We have described five approaches to differential GPS tracking of low earth satellites. The purely geometric strategy is by far the most limited and is included here primarily for illustration and completeness. The general form, in which GPS orbits are adjusted, is capable in principle of near-decimeter performance but requires a prohibitive number of ground sites. The simple form, without GPS adjustment, is practical but limited to meter-level performance. Since it is operationally the simplest technique, geometric tracking may be the method of choice for missions requiring meter-level accuracy.

Fully dynamic tracking, whether with GPS or with another system, can offer decimeter accuracy only so long as dynamic modeling errors are adequately contained. For Topex, the dominant error is in the earth gravity model, and continued success in the current gravity improvement effort will be needed to reach a decimeter. For satellites at lower altitudes, such as NROSS, ERS-1, and EOS, decimeter gravity modeling will present a greater challenge; at the lowest altitudes, where atmospheric drag is dominant, decimeter modeling is far out of reach.

The non-dynamic strategy, with its geometric user solution and dynamic GPS solution, is the first to offer practical subdecimeter accuracy at all altitudes and to dynamically active vehicles. Moreover, with no high-fidelity user models to compute, it is operationally simpler than a dynamic approach. It suffers, however, from a natural sensitivity to weak observing geometry, making it vulnerable to various forms of system degradation which can cause it to fail altogether.

Two rather different extensions of the non-dynamic strategy shore up this weakness by bringing more information to bear. The reduced-dynamic strategy is a sophisticated hybrid bringing together the dynamic and non-dynamic techniques in an optimal combination that can be continuously varied from fully dynamic to non-dynamic. Extensive analysis shows that this strategy must always be equal to or better than either technique separately and that it will enjoy its greatest success when dynamic performance and non-dynamic performance are comparable.

The gravity adjustment strategy is designed to exploit efficiently the information in an ensemble of repeat ground tracks. In general, each arc of the ensemble will reflect a different pattern of GPS satellite formations. The resulting set of position corrections, common to all arcs, will therefore be less sensitive to momentary weaknesses in GPS geometry. (Geographically correlated weaknesses due to ground site distribution will of course persist.) This is a specialized technique which may be of benefit to missions like Topex with a suitable orbit and a delayed processing schedule. For general applications, however, the optimized reduced dynamic strategy appears to be the strongest option. Though somewhat more complex operationally than classical dynamic orbit determination, it offers subdecimeter accuracy to all low orbiters, minimal sensitivity to dynamic and geometric weaknesses, and the versatility to adapt to changing conditions.

Acknowledgment

The authors are grateful to Drs. Stephen M. Lichten and Catherine L. Thornton of JPL, who made numerous valuable contributions.

References

[1] G. H. Born, C. Wunch, and C. A. Yamarone, "TOPEX: Observing the Oceans from Space," *EOS Trans.*, vol. 65, pp. 433–437, July 10, 1984.

[2] G. H. Born, R. H. Stewart, and C. A. Yamarone, "TOPEX–A Spaceborne Ocean Observing System," in *Monitoring Earth's Ocean, Land, and Atmosphere from Space–Sensors, Systems, and Applications*, A. Schnapf, ed., New York: American Institute of Aeronautics and Astronautics, Inc., 1985.

[3] B. D. Tapley and J. C. Ries, "Orbit Determination Requirements for Topex," AAS Paper 87-429, presented at the AAS/AIAA Astrodynamics Specialist Conference, Kalispell, Montana, August 1987.

[4] M. H. Freilich, *The Science Opportunities using the NASA Scatterometer on NROSS*, JPL Publication 84-57, Jet Propulsion Laboratory, Pasadena, California, February 1, 1985.

[5] R. Holdaway, "Assessing Orbit Determination Requirements for ERS-1," AIAA Paper 86-0403, presented at the AIAA 24th Aerospace Sciences Meeting, Reno, Nevada, January 1986.

[6] K. F. Wakker, R. C. A. Zandbergen, and B. A. C. Ambrosius, "Seasat Orbiter Determination Experiments in Preparation for the ERS-1 Altimetry Mission," AAS Paper 87-426, presented at the AAS/AIAA Astrodynamics Specialist Conference, Kalispell, Montana, August 1987.

[7] R. Hartle and A. Tuyahov, "The Earth Observing System," AAS Paper 85-397, presented at the AAS/AIAA Astrodynamics Specialist Conference, Vail, Colorado, August 1985.

[8] B. G. Williams, "Precise Orbit Determination for NASA's Earth Observing System Using GPS," AAS Paper 87-409, presented at the AAS/AIAA Astrodynamics Specialist Conference, Kalispell, Montana, August 1987.

[9] R. J. Milliken and C. J. Zoller, "Principles of Operation of NAVSTAR and System Characteristics," *Navigation*, vol. 25, pp. 95–106, summer 1978.

[10] J. J. Spilker, "GPS Signal Structure and Performance Characteristics," *Navigation*, vol. 25, pp. 121–146, summer 1978.

[11] M. P. Ananda and M. R. Chernick, "High Accuracy Orbit Determination of Near-Earth Satellites using Global Positioning System (GPS)," in *Proceedings, IEEE PLANS '82*, pp. 92–98, 1982.

[12] S. C. Wu and V. J. Ondrasik, "Orbit Determination of Low-Altitude Earth Satellites using GPS RF Doppler," in *Proceedings, IEEE PLANS '82*, pp. 82–91, 1982.

[13] T. P. Yunck, W. G. Melbourne, and C. L. Thornton, "GPS-Based Satellite Tracking System for Precise Positioning," *IEEE Trans. Geosci. and Remote Sensing*, vol. GE-23, no. 4, pp. 450–457, July 1985.

[14] S. M. Lichten, S. C. Wu, J. T. Wu, and T. P. Yunck, "Precise Positioning Capabilities for TOPEX using Differential GPS," AAS Paper 85-401, presented at the AAS/AIAA Astrodynamics Specialist Conference, Vail, Colorado, August 1985.

[15] T. P. Yunck, S. C. Wu, and S. M. Lichten, "A GPS Measurement System for Precise Satellite Tracking and Geodesy," *J. Astronautical Sci.*, vol. 33, no. 4, pp. 367–380, October–December 1985.

[16] W. G. Melbourne and E. S. Davis, "GPS-Based Precision Orbit Determination: A Topex Flight Experiment," AAS Paper 87-430, presented at the AAS/AIAA Astrodynamics Specialist Conference, Kalispell, Montana, August 1987.

[17] T. P. Yunck and S. C. Wu, "Non-Dynamic Decimeter Tracking of Earth Satellites using the Global Positioning System," AIAA Paper 86-0404, presented at the AIAA 24th Aerospace Sciences Meeting, Reno, Nevada, January 1986.

[18] S. C. Wu, S. M. Lichten, and T. P. Yunck, "Gravity Mismodelling on Topex Orbit Determination," AAS Paper 86-2056-CP, presented at the AIAA/AAS Astrodynamics Specialist Conference, Williamsburg, Virginia, August 1986.

[19] S. C. Wu, T. P. Yunck, and C. L. Thornton, "Reduced-Dynamic Technique for Precise Orbit Determination of Low Earth Satellites," AAS Paper 87-410, presented at the AAS/AIAA Astrodynamics Specialist Conference, Kalispell, Montana, August 1987.

[20] F. J. Lerch, S. M. Klosko, R. E. Laubscher, and C. A. Wagner, "Gravity Model Improvement using GEOS 3 (GEM 9 and 10)," *J. Geophys. Res.*, vol. 48, no. B8, pp. 3897–3916, July 1979.

[21] F. J. Lerch, B. H. Putney, C. A. Wagner, and S. M. Klosko, "Goddard Earth Models for Oceanographic Applications (GEM 10B and 10C)," *Marine Geodesy*, vol. 5, no. 2, pp. 145–187, 1981.

[22] F. J. Lerch, S. M. Klosko, G. B. Patel, and C. A. Wagner, "A Gravity Model for Crustal Dynamics (GEM-L2)," *J. Geophys. Res.*, vol. 90, no. B11, pp. 9301–9311, September 1985.

[23] G. Rosborough, "Orbit Error Due to Gravity Model Error," AAS Paper 87-534, presented at the AAS/AIAA Astrodynamics Specialist Conference, Kalispell, Montana, August 1987.

[24] T. Meehan *et al.*, "Rogue: A New High Accuracy Digital GPS Receiver," presented at the International Union of Geodesy and Geophysics Nineteenth General Assembly, Vancouver, Canada, August 1987.

[25] Y. Bock, S. A. Gourevitch, C. C. Counselman III, R. W. King, and R. I. Abbot, "Interferometric Analysis of GPS Phase Observations," in *Proc. Fourth Int. Geodetic Symp. on Satellite Positioning*, Austin, Texas, April 1986.

[26] G. Blewitt, "New Approaches to Carrier Phase Ambiguity Resolution and the Benefits of Simultaneous Fits to Phase and Group Delay Observables," presented at the International Union of Geodesy and Geophysics Nineteenth General Assembly, Vancouver, Canada, August 1987.

[27] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, New York: Academic Press, 1977.

[28] S. C. Wu and C. L. Thornton, "OASIS–A New GPS Covariance and Simulation Analysis Software System," in *Proc. First Int. Symp. on Precise Positioning with GPS*, pp. 337–346, May 1985.

[29] J. T. Wu, "TOPEX Orbit Determination by Solving Gravity Parameters with Multiple Arcs," AAS Paper 85-411, presented at the AAS/AIAA Astrodynamics Specialist Conference, Vail, Colorado, August 1985.

[30] J. T. Wu and S. C. Wu, "TOPEX Orbit Determination by Combining GPS Data from Repeat Orbits," AIAA Paper 86-2216-CP, presented at the AIAA/AAS Astrodynamics Specialist Conference, Williamsburg, Virginia, August 1986.

Table 1. Error model and other assumptions used in covariance analysis

User satellite	Topex (1334 km in altitude)
Number of stations	6 (cf. Fig. 2)
Number of GPS satellites	18
Cutoff elevation	10 degrees at stations 0 degrees at Topex
Data type	P-code pseudorange carrier phase
Data span	2 hours
Data interval	5 minutes
Data noise	5 cm (pseudorange) 0.5 cm (carrier phase)
Carrier phase bias	10 km (adjusted)
Clock bias	3 μsec (adjusted as white process noise)
Topex epoch state	2 km; 2 m/sec (adjusted)
GPS epoch states	2 m; 0.2 mm/sec (adjusted)
Station location	5 cm each component
Zenith troposphere	1 cm
Earth's GM	1 part in 10^8
Gravity	Scaled GEM10–GEML2 (20 X 20 lumped)
Solar pressure	10 percent

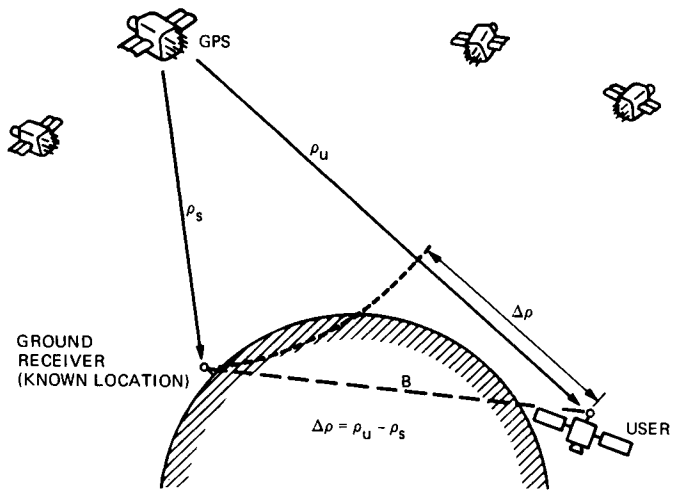


Fig. 1. Differential pseudorange observations to four GPS satellites provide position and time offset with respect to the ground reference point, resulting in substantial cancellation of GPS errors

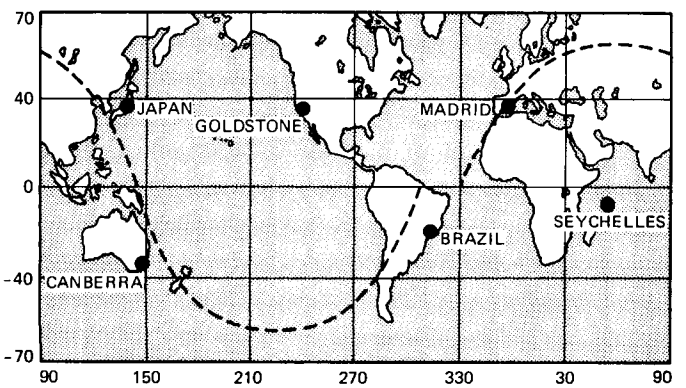


Fig. 2. GPS ground receiver sites used in error studies and one-orbit Topex ground track

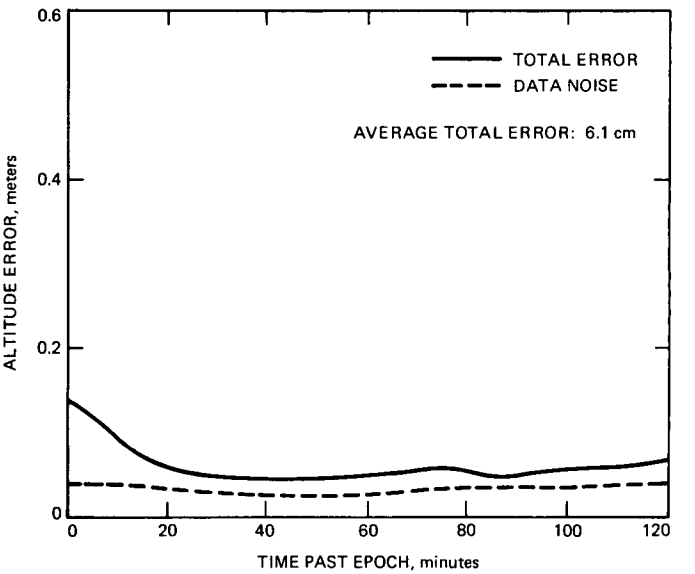


Fig. 3. Predicted Topex altitude error with dynamic differential GPS tracking using an optimistic gravity error model

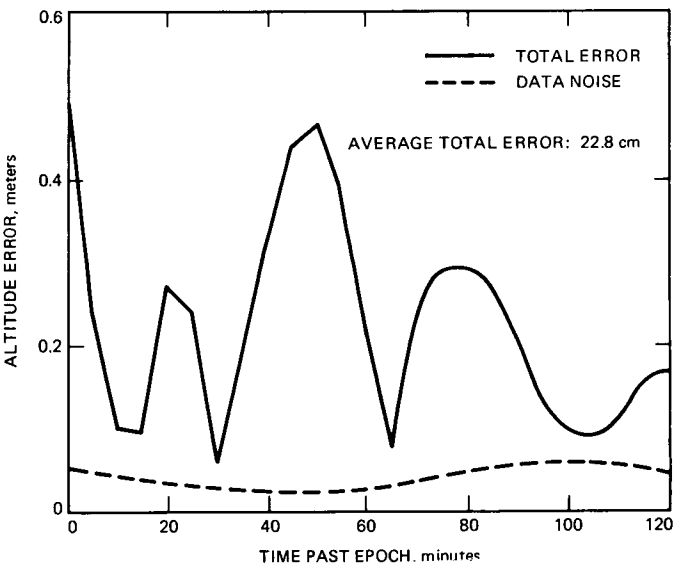


Fig. 4. Predicted Topex altitude error with dynamic differential GPS tracking using a pessimistic gravity error model (c. 1983)

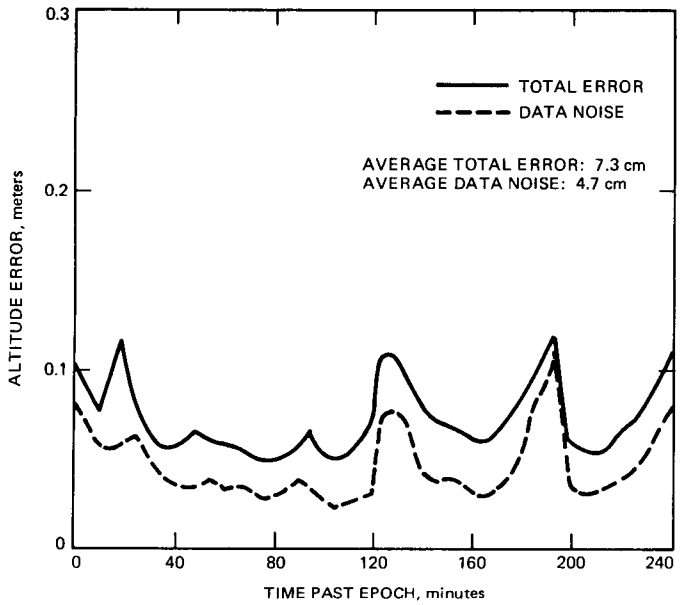


Fig. 5. Predicted Topex altitude error with the non-dynamic strategy using the mixed data type

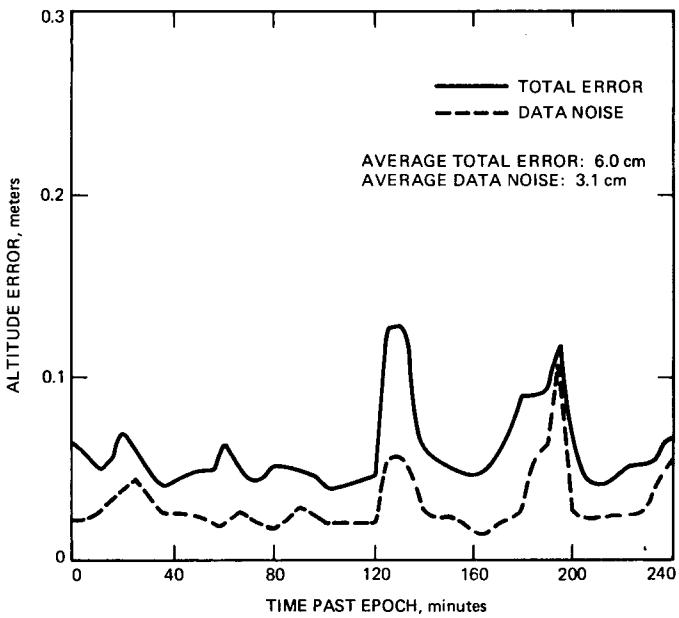


Fig. 6. Predicted Topex altitude error with the non-dynamic strategy using differenced carrier range (τ = 15 minutes)

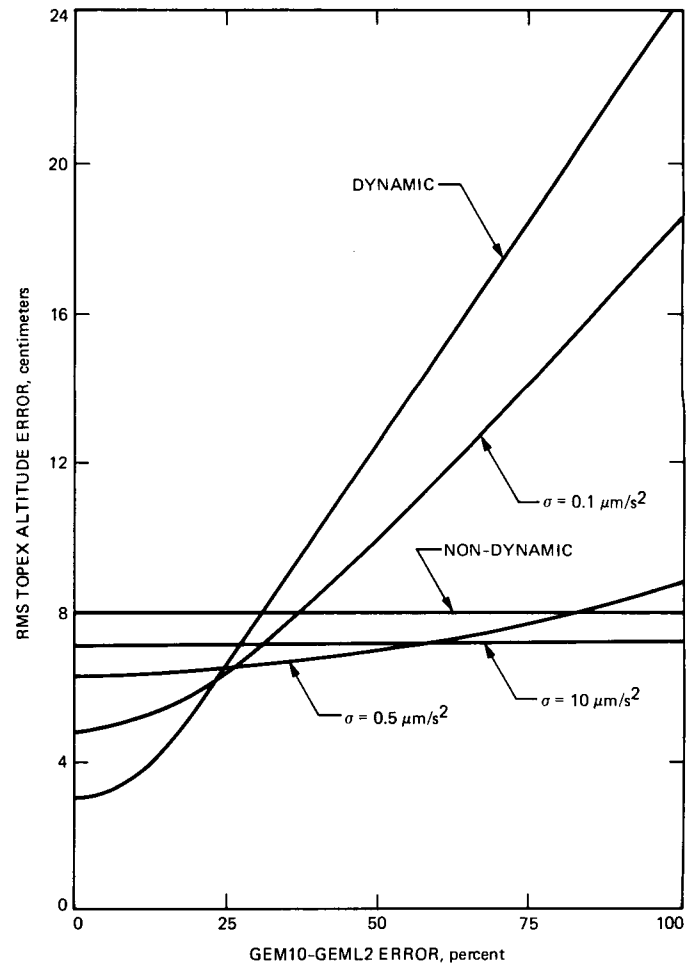


Fig. 7. Predicted Topex altitude error, assuming 6-satellite viewing capacity, using dynamic, non-dynamic, and reduced-dynamic strategies, shown as a function of gravity model quality

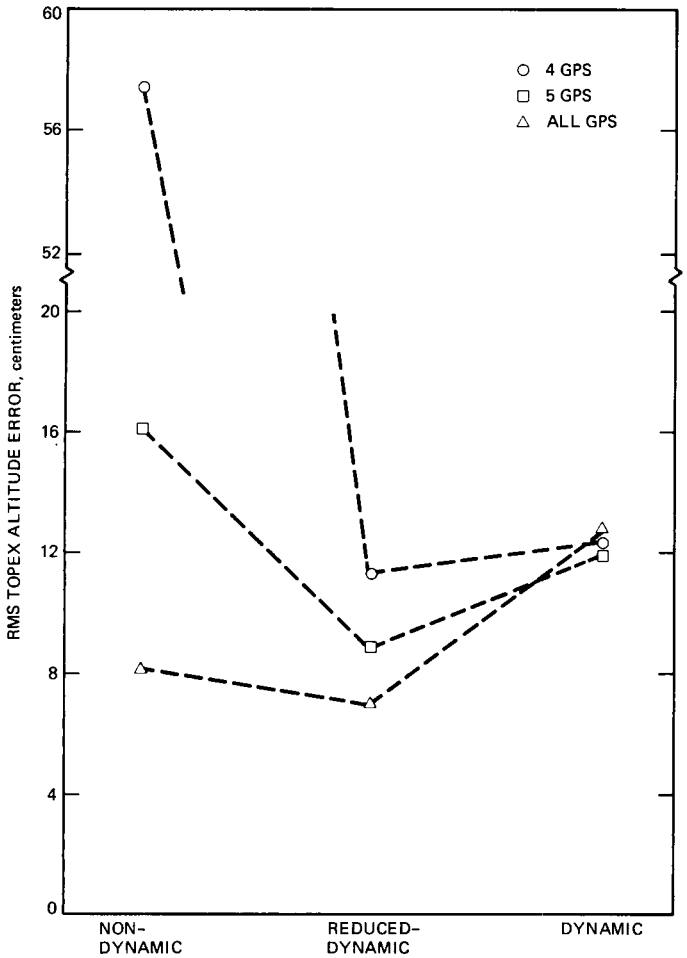


Fig. 8. Predicted Topex altitude error using dynamic, non-dynamic, and reduced-dynamic strategies, shown for three different flight receiver viewing capacities

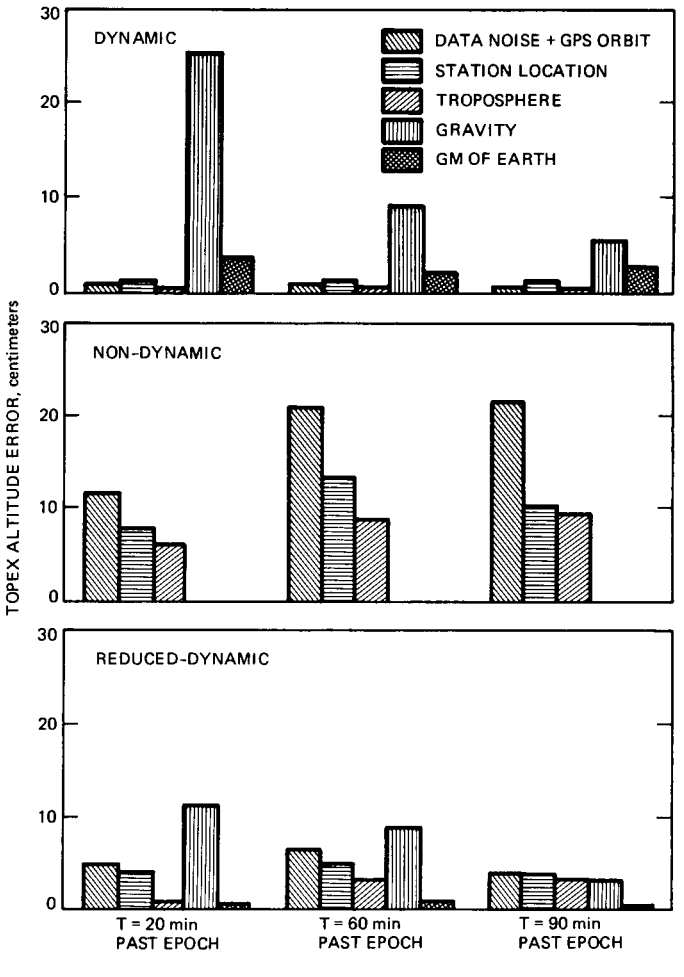


Fig. 10. Breakdown of predicted Topex altitude error at three peak points

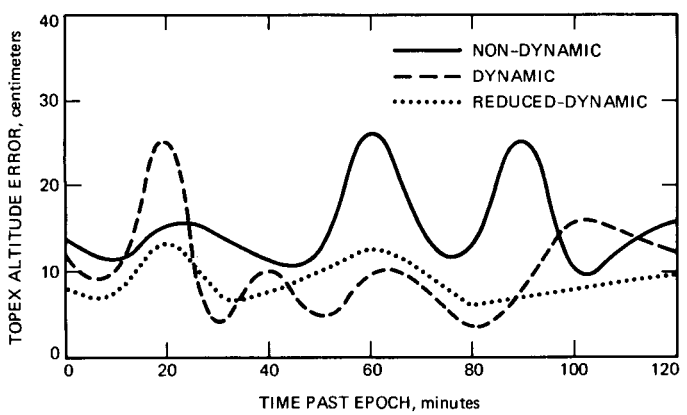


Fig. 9. Detail of predicted Topex altitude error over a full 2-hour data arc for dynamic, non-dynamic, and reduced-dynamic strategies

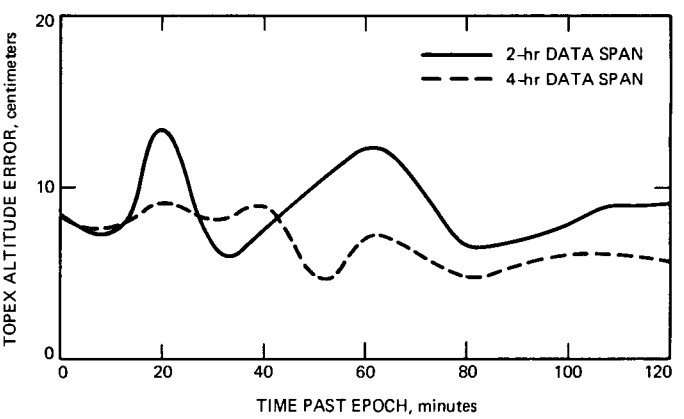


Fig. 11. Detail of predicted Topex altitude error with reduced-dynamic technique for 2-hour and 4-hour data spans

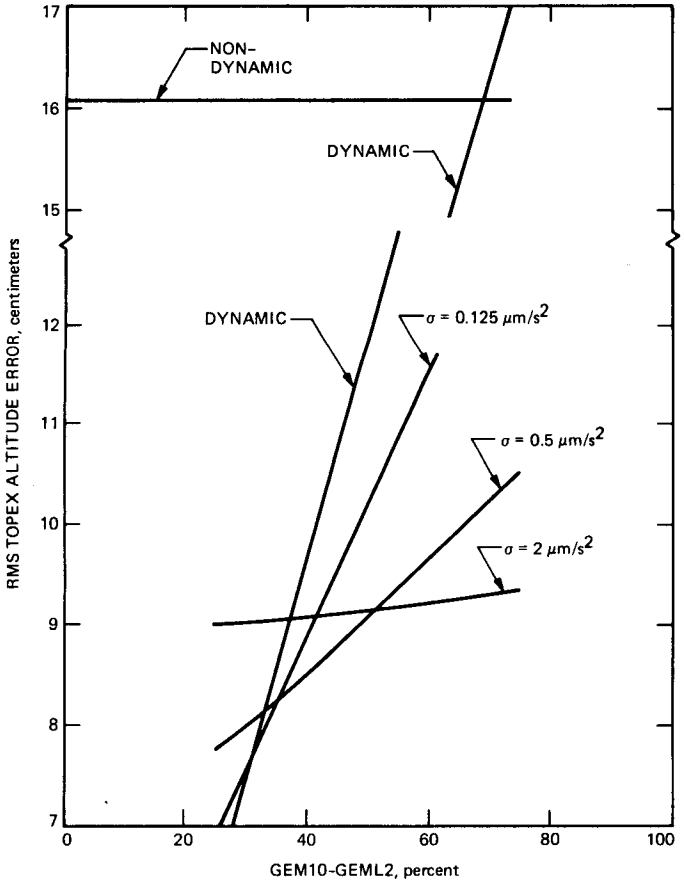


Fig. 12. Predicted Topex altitude error, assuming 5-satellite viewing capacity, using dynamic, non-dynamic, and reduced-dynamic strategies, shown as a function of gravity model quality

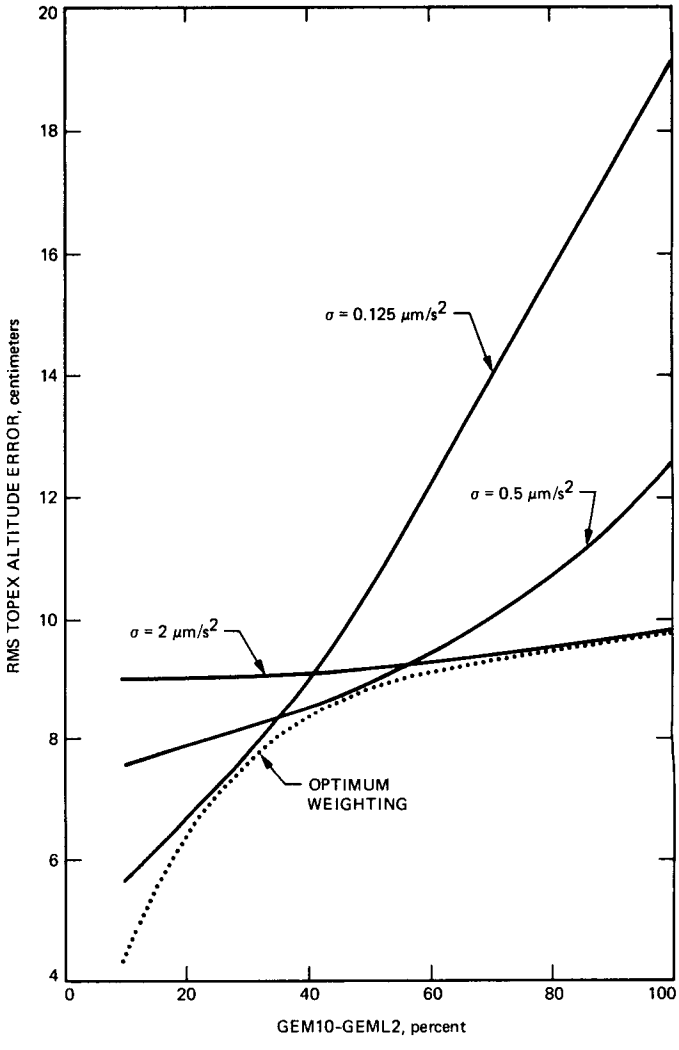


Fig. 13. Predicted Topex altitude error, assuming 5-satellite viewing capacity, with optimal reduced-dynamic weighting (dotted line), shown as a function of gravity model quality

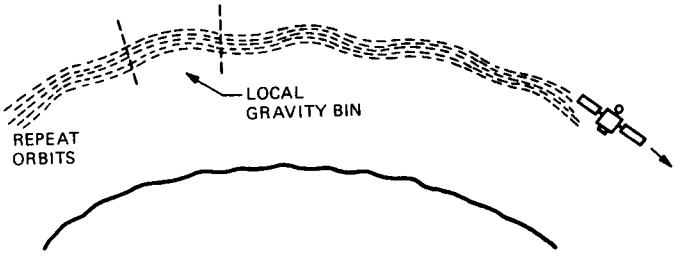


Fig. 14. The gravity adjustment strategy estimates local gravity parameters using data from repeat ground tracks

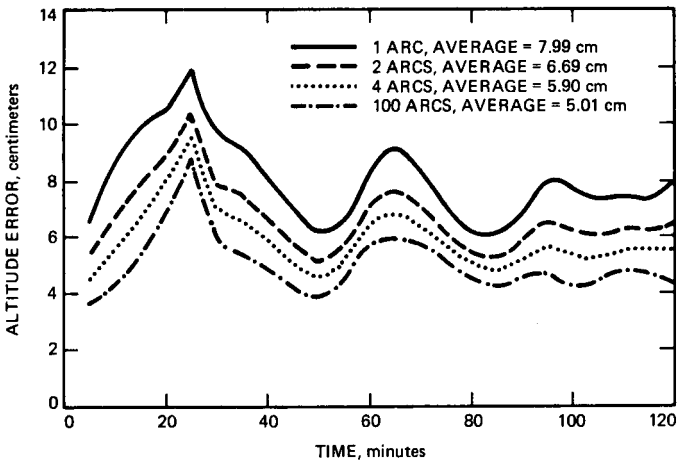


Fig. 15. Predicted Topex altitude error using gravity adjustment strategy, shown for ensembles of repeat arcs numbering from 1 to 100

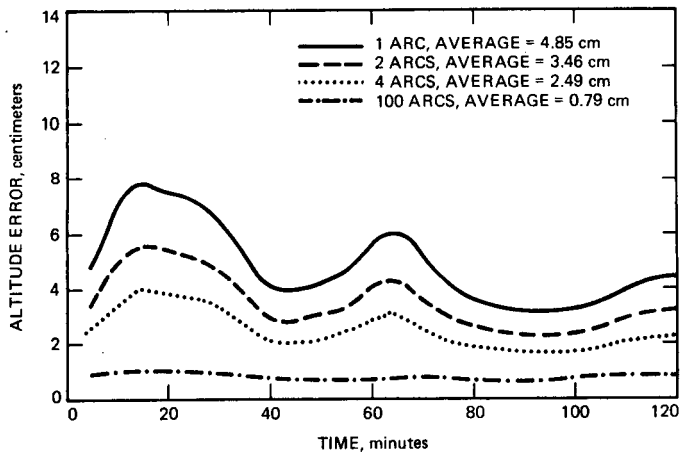


Fig. 16. Data noise contribution to predicted Topex altitude error using the gravity adjustment technique

Short Baseline Phase Delay Interferometry

C. D. Edwards
Tracking Systems and Applications Section

The high precision of the phase delay data type allows angular navigation accuracy on relatively short baselines to compete with the angular accuracy achieved with long baseline group delay measurements. Differential phase delay observations of close quasar pairs on both a 5.9-km baseline (DSS 12–DSS 13) and a 253-km baseline (DSS 13–Owens Valley Radio Observatory) have been performed to study the potential navigational precision and accuracy of short baseline interferometry. As a first step toward demonstration of a connected element system at Goldstone, the DSS 12–DSS 13 baseline was operated coherently, distributing a common frequency reference via a recently installed fiber optic cable. The observed phase delay residuals of about 10 psec or less on both baselines appear to be dominated by short term troposphere fluctuations, and correspond to navigational accuracies of well below 50 nrad for the 253-km baseline. Additional experiments will be required to probe the full range of systematic errors.

I. Introduction

Very long baseline interferometry (VLBI) currently provides high accuracy angular spacecraft navigation for the Deep Space Network (DSN). Relative to the highly stable inertial reference frame defined by extragalactic radio sources, the DSN Block I VLBI system can provide spacecraft angular position measurements with accuracy better than 50 nrad, using intercontinental baselines of lengths up to 10,000 km. The work reported here is an investigation of the extent to which interferometric observations on much shorter baselines can provide competitive results. In terms of ease of operation, reliability, observing bandwidth, and faster access to navigational data, short baseline interferometry offers a number of advantages over intercontinental VLBI. After briefly reviewing interferometry and the different interferometric data types, the advantages of short baselines and connected element interferometry (CEI) will be examined. Recent experimental results for baselines of 5.9 and 253 km will be presented, indicating the current level of position accuracy to be 1–5 mm. Finally, future plans for evaluating the potential of CEI will be considered.

II. Interferometric Phase and Group Delay Data Types

In interferometry, we are essentially interested in measuring the geometric delay representing the difference in arrival times of a signal wavefront at two separate antennas. Letting $\mathbf{B}$ be the baseline between the two stations and $\hat{\mathbf{s}}$ the unit vector in the direction of the radio source, the geometric delay is simply:

$$\tau = \frac{1}{c}\mathbf{B}\cdot\hat{\mathbf{s}}$$

$$= \frac{1}{c}B\cos\theta$$

where c is the speed of light and θ is the angle between the baseline vector and the source direction. This geometric delay, coupled with accurate knowledge of the baseline length and orientation, provides a precise angular position for the radio source. In terms of the delay error $\sigma(\tau)$, the angular position accuracy is:

$$\sigma(\theta) = \frac{c\sigma(\tau)}{B|\sin\theta|}$$

Angular navigation can be improved either by increasing the baseline length or by reducing the delay error.

To measure the delay τ, the signals from the two antennas are brought together and correlated. (See [1], for example.) The primary output of this correlation process is the interferometric phase, which, in the absence of unmodeled errors, is a measurement of the geometric delay in units of the observing wavelength. In terms of the interferometric phase ϕ, the phase delay τ_ϕ is simply:

$$\tau_\phi = \frac{\phi}{\nu_{RF}}$$

where ν_{RF} is the RF observing frequency. The cycle ambiguity associated with the phase determination corresponds to a delay ambiguity of $1/\nu_{RF}$. The fractional number of wavelengths in the geometric delay is measured very accurately, but the integral number is unknown. Therefore, unless *a priori* knowledge of the geometric delay is much better than a cycle of RF phase, and all error sources can be controlled to much better than an RF cycle, the high precision of the phase delay data type cannot be utilized.

In those cases where the *a priori* delay knowledge is not accurate enough to resolve the RF cycle ambiguity, one can instead measure the slope of phase *vs.* frequency to obtain the less precise group delay. In practice, the bandwidth synthesis (BWS) technique is used [2] in which the interferometric phase is observed at two (or more) nearby RF frequencies, yielding a measurement of the group delay τ_{BWS}:

$$\tau_{BWS} = \frac{\phi_1 - \phi_2}{\Delta\nu}$$

where $\Delta\nu$ is the difference of the two observing frequencies. The group delay still has an ambiguity associated with it, of size $1/\Delta\nu$, but since typically $\Delta\nu \ll \nu_{RF}$, the BWS ambiguity is usually much larger than the phase delay ambiguity, and hence much less stringent *a priori* delay knowledge is required to resolve the group delay ambiguity. The increased delay ambiguity, however, is accompanied by a reduced delay precision.

An example will put this in perspective. Consider an interferometric observation with a signal-to-noise ratio SNR = 10. Let the RF observing frequency be ν_{RF} = 8.4 GHz, with observations at two channels separated by $\Delta\nu$ = 40 MHz, typical values for the DSN Block 0 and Block I VLBI systems. Table 1 summarizes the resulting delay precision and ambiguity for both the phase and group delay data types. The phase delay precision of 2 psec is over two orders of magnitude smaller than the 560-psec precision of the group delay. However, the phase delay ambiguity of 120 psec is also over two orders of magnitude smaller than the group delay ambiguity of 25 nsec.

In typical intercontinental VLBI observations, a number of error sources cause the *a priori* model delay error to be on the order of an RF cycle or larger, preventing the use of the high precision phase delay data type. As a result, the group delay has routinely been used to supply delay estimates on long baselines for spacecraft navigation.

On short baselines, however, cancellation of the effects of many of these error sources between the two stations reduces the *a priori* delay uncertainty to well below an RF cycle, enabling phase connection. The dominant delay error sources on intercontinental baselines include uncertainties in the zenith components of the troposphere and ionosphere, as well as angular uncertainties in the *a priori* source position and baseline orientation which produce delay errors proportional to baseline length. For baselines of a few hundred kilometers or less, the delay uncertainty corresponding to angular errors in source position and baseline orientation becomes much less than an RF cycle. Also, partial cancellation of the effects of the troposphere and ionosphere occurs on these shorter baselines, due both to spatial correlations in these media and to the fact that on shorter baselines, both stations are observing at roughly the same elevation angle. On sufficiently short baselines, the delay error budget will be dominated by baseline length-independent errors such as instrumental stability and antenna flexure, and by rapid fluctuations in the density of water vapor over each antenna, on a scale size smaller than the antenna separation.

Reduction of delay errors can also be achieved by limiting observations to sources within a small region of the sky. Delay errors which vary slowly over the sky will cancel to a large extent between observations of angularly close sources. (This strategy is used in Delta Differential One-way Range [Delta-DOR] navigation observations, in which a spacecraft's angular position is measured with respect to a nearby quasar.) By combining close source separations with short baselines, delay errors can be substantially reduced through cancellation, fur-

ther enabling phase ambiguity resolution. Of the two phase connection experiments reported here, one corresponds to a baseline length of 5.7 km with source separations of 10–20 degrees on the sky, while the other corresponds to a 253-km baseline with roughly 2-degree source separations.

III. Advantages of Short Baseline Interferometry

The use of intercontinental baselines is motivated by the fact that, for a given interferometric delay accuracy, the resulting angular accuracy is inversely proportional to baseline length. As we have seen, however, use of the more precise phase data type on shorter baselines can provide results competitive with group delay accuracy on longer baselines. In addition, a number of operational advantages result from observing on shorter baselines.

For intercontinental baselines, only a limited portion of the sky is simultaneously visible from both antennas. For example, on the Goldstone–Canberra DSN baseline, with an elevation angle cutoff of 6 deg, the solid angle visible from both complexes at any given time is only 32 percent of 2π sr. On the other hand, for observations within a single DSN complex, with the same 6-deg elevation cutoff the instantaneous mutual visibility approaches 90 percent of 2π sr. This would significantly relax scheduling constraints for tracking during critical mission periods.

A related problem with intercontinental baselines is that observations must often be made with one or both stations at very low elevation angles, increasing the effects of the troposphere. As an example, consider intracomplex Goldstone observations of a source at –23-deg declination. At transit, the source will be at an elevation angle of 31.7 deg, corresponding to a total pathlength through the atmosphere, summed over both antennas, of 3.8 atmospheres. For the Goldstone–Canberra intercontinental baseline, the minimum total pathlength is only slightly larger, at 4.2 atmospheres, but for the Goldstone–Madrid baseline, the minimum total pathlength is 15.0 atmospheres. For those scenarios in which troposphere errors are a dominant part of the error budget, the capability of observing at higher elevation angles could significantly improve navigational accuracy.

The possibility of using the phase delay observable on shorter baselines also reduces the impact of instrumental phase uncertainties in the observing bandpass. For a differential spacecraft–quasar observation, the phase of the spacecraft tone will be affected by instrumental phase errors at the ground stations only at the tone frequency in the observing bandpass, whereas the phase of the broadband quasar signal is subjected to the average of the instrumental phase error across the observing bandpass. Any phase ripple across the bandpass will not cancel in the spacecraft–quasar difference and will corrupt the final navigational observable. This can produce a significant delay error in the group delay, equal to the size of the phase error divided by the spanned bandwidth. For example, with an uncalibrated phase ripple of 2 deg across the observing bandpass and a spanned bandwidth of 40 MHz, the resulting error in the group delay would be $\sqrt{2} \times (0.006 \text{ cyc})/(40 \text{ MHz})$, or about 200 psec. For the phase delay data type, however, the effect is just the phase error divided by the RF observing frequency. Thus at 8.4 GHz, the 2-deg phase ripple would only induce an error of 0.7 psec.

For baseline lengths of a few tens of kilometers or less, the observing stations can be directly connected by fiber optic cables, allowing the observed signals to be transported to a central correlator facility and processed in real time. In addition to decreasing the turnaround time for navigational data, real-time cross-correlation would improve reliability by providing verification of the data integrity, revealing any experimental configuration errors at the time of observation rather than hours or days later. The fiber optic links also enable distribution of a common frequency reference to the observing stations, eliminating the requirement for separate clocks and the need to solve for a clock rate offset between stations.

An additional advantage of real-time correlation is that data would not have to be recorded on tape, but would rather be routed directly to the correlator for processing. Eliminating the need for data recording would increase efficiency, eliminate costs for tapes and tape shipping, and greatly reduce the manpower required for correlation processing. With the burden of actually recording the observed signals removed, much larger observation bandpasses and consequent data rates could be supported, leading to improved signal-to-noise for the broadband quasar observations. This improved sensitivity could allow observation of weaker radio sources, increasing the density of the source catalog, and could shorten observing times for current radio sources.

Finally, it should be noted that the phase delay spacecraft measurement is a narrow band technique. The phase delay is determined from data at a single frequency, as opposed to the group delay, which requires observations at several frequencies. Just a single carrier wave spacecraft tone is required, rather than a set of tones with tens or hundreds of MHz separation. Charged particle calibration would still necessitate dual-band data, but again, only one tone at each band would be required, rather than several. Quasar observations will still require finite bandwidth, since the quasar cross-correlation SNR scales inversely with the square root of the recorded bandwidth. However, the large spanned bandwidths used in BWS are not

required. The data presented in this article were recorded with single 2-MHz bandpasses at 2.3 and 8.4 GHz.

IV. Goldstone Intracomplex Phase Delay Observations

A. Initial Tests of Fiber Optic Link

As a first step toward CEI at Goldstone, and to support a number of other intracomplex communication needs, four multi-mode and two single-mode fiber optic cables were installed between DSS 12 and DSS 13 in 1986. The single-mode fibers each offer over 1 GHz of bandwidth, and could eventually be used to transmit data to a central correlator facility. Currently the single-mode fiber is being used not for data transfer, but rather for transfer of a frequency reference from DSS 13 to DSS 12, enabling both stations to be operated coherently.

The first interferometry experiments using the fiber optic link between DSS 12 and DSS 13 were completed in August and September 1986. The fiber optic link was used to reference the local oscillators at both stations to the hydrogen maser frequency reference at DSS 13. The goal of these experiments was to verify the coherent operation of the two stations by successfully finding interferometric fringes, and secondly, to characterize the stability of the link by examining the phase residuals for individual quasar scans. Observations of a number of bright radio sources were made at 2.3 and 8.4 GHz, and a 2-MHz bandwidth at each frequency was translated to baseband, single-bit sampled, time-tagged, and recorded on videotape, using the Block 0 VLBI system. The experimental configuration is summarized in Fig. 1.

Figure 2 shows the Allan standard deviation of the phase residuals for a single 10-minute scan of the bright source 3C 84. At τ = 64 sec, the Allan standard deviation $\sigma_y(\tau) = 4.5 \times 10^{-14}$. Also indicated in the figure is the expected range of stability due to troposphere fluctuations, based on a calculation using the stochastic model of Treuhaft and Lanyi [3]. Instabilities in the VLBI instrumentation may also contribute at about this level or slightly lower.[1] The agreement indicates that wet troposphere fluctuations and perhaps station instrumental instabilities can account for the observed phase fluctuations, suggesting that the stability of the fiber optic link for frequency transfer is below this level. Independent measurements of the round trip link stability by members of the Time and Frequency Systems Research Group at JPL yielded an Allan standard deviation of 1.5×10^{-15} at τ = 1000 sec, and below 10^{-14} at τ = 64 sec [4], again indicating that the fiber optic link is more stable than the expected level of wet troposphere fluctuations.

B. Phase Connection Results for Clusters of Sources

To demonstrate phase connection on short baselines, interferometric observations of a number of extragalactic radio sources were made between DSS 12 and DSS 13 on March 21, 1987. Data were recorded at 2.3 and 8.4 GHz using the same Block 0 configuration as described above. Most of the observations were 3 minutes long and were grouped into clusters of 8–10 scans. The goal of this experiment was to determine the reliability of phase ambiguity resolution, and quantify limiting error sources, by examining the final RMS scatter of phase delay residuals within each cluster.

Each cluster consisted of repeated observations of two or three angularly close radio sources. Separation angles for sources within a cluster ranged from 10 to 20 deg, representative of typical spacecraft–quasar angular separations in current Delta-DOR navigation observations. As mentioned earlier, limiting observations to angularly close sources leads to significant cancellation of many geometric and propagation media error sources, thereby decreasing the relative *a priori* model delay errors between sources and increasing the likelihood of successful phase connection.

The data were correlated off-line at the Caltech–JPL Block 0 Correlator, yielding phase measurements for each scan at both 2.3 and 8.4 GHz. Calibration tones embedded in the recorded bandpasses were also extracted, and the resulting instrumental phase variations at each station were removed from the phases for the radio sources. Based on the best available *a priori* knowledge of the positions of the radio sources and the station locations for DSS 12 and DSS 13, a model delay was calculated for each scan, and the residual delay difference between the observed phase delay and this model delay was calculated. These relative phases were then examined to correct the RF cycle ambiguities for these scans. The phase ambiguity of each scan was adjusted so that its residual delay was within half an ambiguity of the residual delay for the preceding scan. Finally, S/X ionosphere calibration was performed by forming the linear combination of 2.3- and 8.4-GHz observations which eliminated the dispersive effects of the ionosphere.

At this point in the analysis, the dominant uncertainty in the phase delay model is an unknown relative phase between the local oscillator chains at the two stations, attributable to phase uncertainties in the frequency transfer between stations and uncalibrated portions of the signal chain (e.g., the path between the antenna feed horn and the injection port for the phase calibration tones). This phase uncertainty manifests

[1] C. D. Edwards, IOM 335.4-558 to Tracking Systems and Applications Section (internal document), Jet Propulsion Laboratory, Pasadena, California, January 1986.

itself as a clock epoch offset between the two stations. Standard VLBI analysis routinely includes a clock term in the delay model to characterize the different clock behavior at the two stations. In addition to a constant clock epoch offset, one must typically include a linear term and sometimes a quadratic term to characterize frequency offsets and drifts between the clocks at the two sites. For this experiment, however, both stations were operated coherently, meaning that only the constant clock epoch term was required.

A weighted least-squares fit was used to adjust the final delay model to the observed phase delays. The DSS 12 station location and the differential zenith troposphere between stations were adjusted. In addition, a constant clock offset was estimated for each cluster of scans. The estimation of a separate clock epoch for each cluster means that the phase is only being connected between sources within a cluster. This makes the solution less sensitive to errors in station location, or errors which vary slowly over the sky, such as antenna deformation. The multiple clock epoch estimates will also tend to compensate for any slow, long term uncalibrated phase drifts between the instrumentation at the two stations. The shift in the DSS 12 station location, relative to the *a priori* location, was $\Delta x = -10 \pm 6$ mm, $\Delta y = 8 \pm 5$ mm, and $\Delta z = 1 \pm 7$ mm, while the differential troposphere adjustment was 0 ± 1 mm.

The final S/X delay residuals are plotted in Fig. 3 against UT of the observation. The clusters are separated by dashed vertical lines, with a size of 120 psec, or one cycle of phase at 8.4 GHz. The scatter of the data is much smaller than this ambiguity size, indicating the reliability of the phase ambiguity resolution. The statistical errors on the individual data points, based on signal-to-noise considerations, were below 1 psec for nearly all the scans. The actual size of the limiting systematic errors is reflected in the scatter of the repeated observations within a cluster. The RMS scatter of the residual delays for each cluster ranges from 2.8 psec up to 15.7 psec. It is interesting to note that the largest scatter is obtained for the lowest elevation data: the observations of DW 1335-12 and OP-192 in the third data cluster, for which the RMS delay scatter is 15.7 psec, were made at elevations of 8–20 deg. Aside from one scan in the final cluster, all other observations were made at elevations of 35–75 deg. The combined RMS delay residual for all 37 observations is 9.1 psec. While the sources of the limiting systematic errors are not fully known, the model of [3] suggests that troposphere fluctuations can account for much of the observed scatter. Table 2 summarizes these results. It should be kept in mind that the only parameters estimated in fitting the data were DSS 12 station location, differential troposphere delay, and the clock epoch offsets for each cluster of observations; the rest of the delay model, including the radio source positions, were fixed at their best *a priori* values.

(As a further test of the robustness of the phase connection solution, an additional solution was made adjusting only the clock epoch parameters; the station location and troposphere values were fixed at their *a priori* values. Similar results were obtained, with the combined RMS delay scatter increasing only slightly to 10.7 psec.)

V. Phase Connection Between Goldstone and Owens Valley

As encouraging as these results on the DSS 12–DSS 13 baseline are, they do not translate into very accurate angular navigation. Even with an optimal source–baseline geometry, 10-psec delay accuracy on a 5.9-km baseline only provides roughly 600-nrad angular position accuracy. Competing with intercontinental capabilities will require improved delay accuracy, longer baselines, or a combination of both.

To investigate the potential for phase connection on longer baselines, VLBI observations were made on the 253-km baseline between DSS 13 and the 40-meter telescope at Owens Valley Radio Observatory (OVRO) on May 20, 1986. The primary difference between these observations and the DSS 12–DSS 13 observations involves clocks: for the DSS 13–OVRO experiment, each station was referenced to its own hydrogen maser frequency standard, as opposed to the DSS 12–DSS 13 experiment for which the Goldstone fiber optic link was used to operate both stations coherently. Other than that, most experimental details were the same. Block 0 VLBI recording systems were used and correlation was again performed off-line at the Caltech/JPL Block 0 Correlator.

Observations were made of two pairs of angularly close radio sources: GC 1633+38 and 3C 345, which are separated by 2.25 deg, and 3C 371 and 1749+701, with a separation angle of 1.59 deg. These separation angles are considerably smaller than the 10–20 deg separations of the sources within the clusters used in the DSS 12–DSS 13 observations. As a first attempt at phase connection on this baseline, the smaller separation angles were chosen to help compensate for geometric error sources which depend on both baseline length and source separation. (For example, according to the model of [3], the RMS differential zenith troposphere delay on a 5.9-km baseline is only about 0.5 cm; for 253 km it rises to over 2 cm. Reducing source separations will reduce the differential effect of this error source.) Future experiments will attempt phase connection for larger source separations, using water vapor radiometers to calibrate the troposphere at each station.

As before, residual phase delays were formed by differencing the observed interferometric phase delay and a model

delay based on *a priori* values of station locations, source positions, earth orientation, zenith troposphere delays, etc. The dominant effect in these phase delay residuals is the clock behavior. However, because independent clocks were used at DSS 13 and OVRO, the clock modeling is more complicated for this experiment than for the DSS 12–DSS 13 fiber optic link experiment. The separate clocks will induce an unknown phase rate (of order 1 psec/sec) corresponding to a frequency offset between the two clocks, in addition to stochastic instabilities in each clock. To make use of the observed phase delays for the individual scans, we would have to model the relative behavior of the two clocks as a piecewise linear or quadratic function over the full experiment. To simplify the analysis and reduce the need to accurately model the clock behavior over long time periods, individual observations were combined to form differential observables for which the clock behavior drops out, as described below.

For a given quasar pair, a sequential triplet of observations, first of source A, then source B, and back to source A, was selected. Individual scans were 5 minutes in length, and each A-B-A sequence lasted from 16 to 32 minutes, depending on details of the observing schedule. The relative phases for the two observations of source A were adjusted based on the observed phase rates for those scans. (More explicitly, for the two observations of source A, a predicted phase difference was obtained based on the average of the observed phase rates and the time difference between observations. The relative phase ambiguities of the two source A observations were than adjusted to bring the actual phase difference to within half an ambiguity of this predicted phase difference.) These two phase delay observations of source A were then used to solve for a linear clock, and this linear clock was interpolated to the time of observation for source B. The phase ambiguity of source B was then adjusted to be within half a cycle of this interpolated clock term. Finally, the differential observable was then formed by subtracting the interpolated clock term from the observed phase delay residual for source B. It is worth noting that, in addition to a linear clock term, any other error sources which are constant or vary linearly with time over the A-B-A observing sequence will also be removed by this technique.

Four such differential observables were formed for each of the two quasar pairs, representing a total of 24 individual radio source observations. The residual differential delays are plotted in Fig. 4 as a function of observation epoch, and the scale of the 8.4-GHz delay ambiguity is indicated. The RMS residual scatter of the four observations of the differential delay between GC 1633+38/3C 345 was 8.9 psec. For 3C 371/1749+701, the RMS scatter was 3.1 psec. The observed RMS scatter is consistent with the expected level of temporal troposphere fluctuations on the 1000-sec time scale associated with individual A-B-A sequences, according to the model of [3].

For the small source separations used here, and given the fact that the observations were all above 45-deg elevation, the static troposphere error due to uncertainty in the zenith delay at each site is expected to be smaller than this fluctuating component. This suggests that phase connection may be possible on this baseline for much larger source separations, particularly if the length of the observing sequences can be shortened. Once again, the fact that the RMS of the residuals is much smaller than the delay ambiguity indicates that the phase has been correctly connected between the source pairs within each linear combination.

For this experiment, no parameters have been estimated. Forming the differential observables eliminates the need to solve for any clock parameters. The close proximity of the sources within each pair makes the differential delay quite insensitive to errors in station location or differential troposphere delay. With 2-deg source separation, for example, a 1-cm station location error would only affect the differential delay at the level of 1.2 psec or less, depending on the relative baseline–source geometry. Accurate *a priori* station locations and zenith troposphere delays would become more important for larger source separations. Table 3 summarizes the results of the DSS 13–OVRO data.

VI. Discussion and Future Plans

It is encouraging that the delay residuals for the 253-km baseline are of roughly the same size as for the 5.9-km baseline. Although the angular source separations were quite small for the DSS 13–OVRO data, the baseline was over 40 times larger. The delay uncertainties of 3–15 psec correspond to a path length uncertainty of 1–5 mm. On the DSS 13–OVRO baseline a 1-cm delay error would correspond to about a 50-nrad angular accuracy for an orientation of 45 deg between the baseline vector and source direction. Based on the very limited amount of data presented here, it is premature to try to estimate a current angular accuracy; nevertheless, the data suggest that systematic errors on differential observations of nearby sources may be well under 1 cm, even for baseline lengths of several hundred kilometers. If differential errors could be reduced to the 1-mm (3-psec) level, differential position accuracies of 50–100 nrad could be obtained on the 21.6-km Goldstone intracomplex baseline between DSS 13 and DSS 14.

Further short baseline phase delay experiments will be required to isolate the dominant error sources and understand their dependence on baseline length, source separation, and observing strategy. Repeated observations of source pairs over many experiments will more fully sample the range of possible systematic errors, providing a more reliable determination of the potential navigational accuracy.

The high precision of the phase delay data type should help in identifying and characterizing the various systematic errors. It may be possible to parameterize and remove certain deterministic errors, such as gravity-induced antenna deformations, if the high precision phase data can reveal the signature of the error in the delay residuals. Stochastic errors such as instrumental phase drifts or wet troposphere fluctuations will require improved calibration techniques. Reliable instrumental phase calibrations and line-of-sight water vapor radiometry will almost certainly be required to reduce differential errors to the 1-mm level for source separations of 10–20 deg.

VII. Summary

Current intercontinental spacecraft navigation makes use of the group data type. The considerably more accurate phase delay data type is not used due to problems resolving the associated cycle ambiguities. On shorter baselines of tens or even hundreds of kilometers, however, phase connection should be possible for reasonable source separations. Using the higher accuracy phase delay data type, short baseline navigation accuracy is competitive with long baseline group delay results. In addition, there are a number of advantages for short baseline and connected element observations over intercontinental VLBI in terms of reliability, efficiency, and visibility.

Frequency transfer over a fiber optic cable enables coherent operation of the 5.9-km baseline between DSS 12 and DSS 13. Stability analysis of phase residuals for single source observations on this baseline indicates that tropospheric fluctuations dominate any instabilities in the fiber optic link frequency transfer. Repeated measurements of clusters of angularly close quasars, with separation angles of 10–20 deg, yield an RMS phase delay scatter of 10 psec, or 3 mm.

Successful phase connection was also achieved on the 253-km baseline between DSS 13 and OVRO, albeit for closer source separation angles of about 2 deg. Due to the use of two independent frequency standards on this longer baseline, linear combinations were formed for observations of neighboring radio sources, eliminating the effect of clock epoch and clock rate offsets. The RMS scatter of the differential measurements was below 10 psec. A delay accuracy of 30 psec, or 1 cm, on this baseline would yield a differential position accuracy of 50 nrad.

A large number of experiments will need to be carried out in the future to better understand the phase delay error budget, and the dependence of various errors on baseline length and source separation.

References

[1] J. B. Thomas, *An Analysis of Radio Interferometry with the Block 0 System*, JPL Publication 81-49, Jet Propulsion Laboratory, Pasadena, California, December 15, 1981.

[2] A. E. E. Rogers, "Very Long Baseline Interferometry with Large Effective Bandwidth for Phase-delay Measurements," *Radio Sci.*, vol. 5, pp. 1239–1247, October 1970.

[3] R. N. Treuhaft and G. E. Lanyi, "The Effect of the Dynamic Wet Troposphere on Radio Interferometric Measurements," *Radio Sci.*, vol. 22, pp. 251–265, March–April 1987.

[4] G. Lutes and A. Kirk, "Reference Frequency Transmission Over Optical Fibers," *TDA Progress Report 42-87*, vol. July–September, Jet Propulsion Laboratory, Pasadena, California, pp. 1–9, November 15, 1986.

Table 1. Properties of phase delay vs. group delay

	Phase delay	Group delay (BWS)
τ	$\frac{\phi}{\nu_{RF}}$	$\frac{\phi_1 - \phi_2}{\nu_1 - \nu_2}$
σ_τ	$\frac{\sigma_\phi}{\nu_{RF}}$	$\frac{\sqrt{2}\sigma_\phi}{\nu_1 - \nu_2}$
	(=2 psec)	(=560 psec)
Ambiguity	$\frac{1}{\nu_{RF}}$	$\frac{1}{\nu_1 - \nu_2}$
	(=120 psec)	(=25 × 10^3 psec)

Numerical values are for the case SNR = 10, ν_{RF} = 8.4 GHz, and $\nu_1 - \nu_2$ = 40 MHz.

Table 2. Intracomplex Goldstone phase delay results

Section	Sources	Separation angle	Number of observations	RMS delay
1	3C 84 NRAO 140 OE 400	6.4–16.2 deg	10	3.1 psec
2	OJ 287 P 0735+17	18.3 deg	8	7.9
3	DW 1335–12 OP–192	5.4 deg	8	15.7
4	OJ 287 P 0735+17	18.3 deg	8	7.6
5	P 1144–379 DW 1335–12 OP–192	5.4–36.5 deg	3	2.8
		Total	37	9.1 psec

DSS 12–DSS 13 baseline length: 5.9 km.
Fiber optic link used to transfer DSS 13 H-maser reference to DSS 12.

Table 3. DSS 13–OVRO phase delay results

Section	Sources	Separation angle	Number of observations	RMS delay
1	GC 1633+38 3C 345	2.25 deg	4 differential	8.9 psec
2	3C 371 1749+701	1.59 deg	4 differential	3.1
		Total	8 differential	6.7 psec

DSS 13–OVRO baseline length: 253 km.

Differential delay observable is formed from A-B-A sequences of single source observations, removing signature of linear clock.

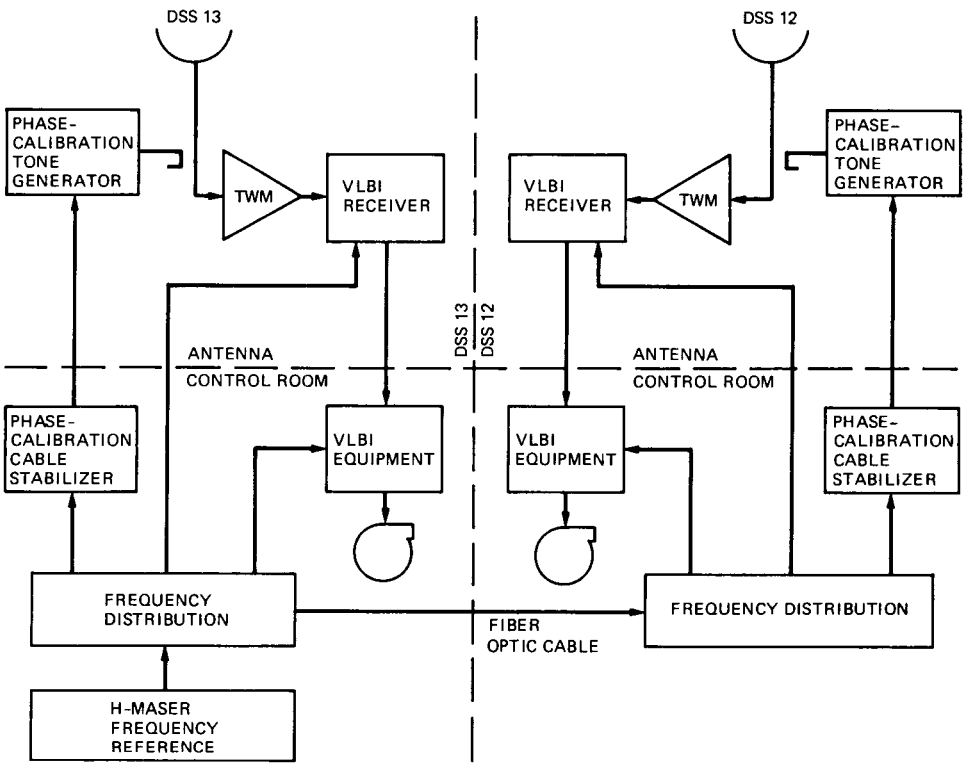


Fig. 1. Experimental configuration for phase delay experiments at Goldstone DSCC. A fiber optic cable transfers a 100-MHz frequency reference from the hydrogen maser at DSS 13 to DSS 12, establishing coherence between the local oscillators at the two stations. Otherwise, the configuration is similar to standard Block 0 VLBI experiments, with data recorded at each station. A full CEI implementation would eliminate the need for data recording, using the fiber optic system for transmission of the observed signals to a common site for real-time correlation.

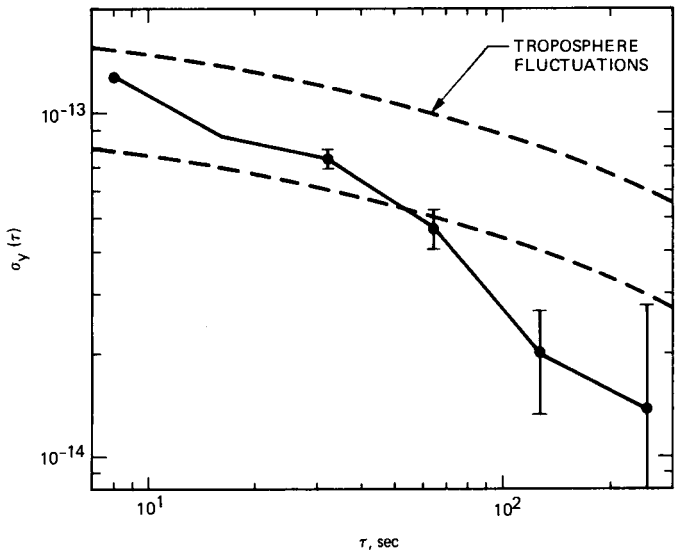


Fig. 2. Allan standard deviation σ_y vs. sample time τ for phase residuals of a 10-minute observation of the bright radio source 3C 84 on the DSS 12-DSS 13 baseline. The dashed lines indicate the range of expected stability due solely to the dynamic wet troposphere, based on a turbulence theory model of water vapor fluctuations. The agreement suggests that such troposphere fluctuations dominate the observed stability, and that the frequency transfer over the fiber optic link is below this level.

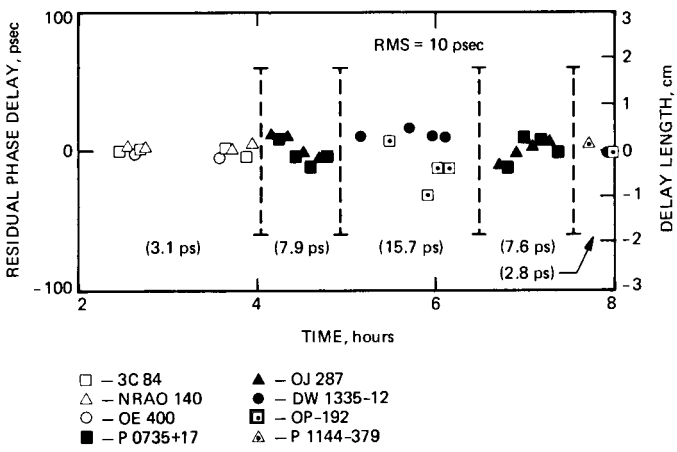


Fig. 3. Phase delay residuals for clusters of 2-3 angularly close radio sources on the DSS 12-DSS 13 baseline. Each source is indicated by a distinct symbol. The vertical dashed lines delimit series of repeated observations for each cluster, and have a height equal to the delay ambiguity at the 8.4-GHz observation frequency. Separate clock offsets were estimated for each series. The RMS scatter for each series is indicated, ranging from 2.8-15.7 psec. The total RMS for all observations is 10 psec.

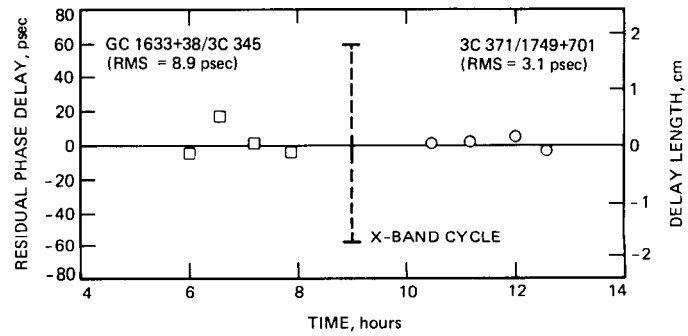


Fig. 4. Differential phase delay residuals for repeated observations of two close quasar pairs, on the 253-km baseline between DSS 13 and the OVRO 40-m antenna. As described in the text, linear combinations are formed for A-B-A observation sequences of each quasar pair to eliminate any linear clock behavior. The total RMS scatter for all eight differential measurements (formed from 24 single source measurements) is 6.7 psec.

Non-Linearity in Measurement Systems: Evaluation Method and Application to Microwave Radiometers

C. T. Stelzried
Office of Telecommunications and Data Acquisition

A simple method for determination and correction of non-linearity in measurement systems is presented. The technique is applicable to a wide range of measuring systems. The basic concept, an analysis, and a sample application are given. Non-linearity of the Goldstone DSS 13 low noise HEMT 2.3-GHz (S-band) radiometer system results in noise temperature measurement errors. These errors are successfully corrected with this method.

I. Introduction

The Deep Space Network (DSN) requires a large number of measurement systems for engineering and science data. For example, microwave Noise Adding Radiometers (NARs) are used for system noise temperature measurements [1]-[3] with applications for the determination of radio "star" flux density for both antenna efficiency calibrations and scientific value.

Engineering design and test procedures are used to obtain linear system measurement performance. A linear system provides equal changes in output reading for equal changes at the system input, independent of signal level. If a noise diode is turned on and off while the microwave radiometer system is switched to the antenna (cold sky) and then to the ambient load, a linear system will give equal changes in output readings. A non-linear system does not.

Linearity of the DSN NAR, a subassembly of the Precision Power Monitor Assembly, is checked with this technique.[1] If the measured noise temperature of a small noise diode, while the system is sequentially switched to the antenna and then to the ambient load, differs by less than 2 percent, the system is considered to be sufficiently linear for use as an NAR.

The purpose of this article is to evaluate an algorithm [4] useful for a wide variety of measurement systems to determine and correct for this type of error.

System non-linearities are only a single example of many possible measurement error sources. Other sources of error, not discussed in this analysis, include such items as incorrectly calibrated noise standards and nonrepeatable front end switches for microwave radiometer systems. Each separate error source should be systematically analyzed to obtain the required overall measurement accuracy. The following analysis provides a method for including non-linearity errors in this list.

II. Theory

A. Linear Analysis

For a linear system, assume an input unknown, T, to be determined in terms of constants and an output reading R

[1] "Precision Power Monitor Assembly," TM 03115A, JPL internal document, pp. 1-3 and 1-4, June 1, 1986.

$$T = A + BR \tag{1}$$

where

T = system input parameter (for example, temperature, K)

A, B = constants, to be evaluated, defining a particular system

R = system output reading (arbitrary units: K, mW, etc.)

For a microwave receiving system,

$T = T_{op}$ = system operating noise temperature for a receiving system [5], K

$= T_i + T_e$

T_i = source temperature, K

T_e = receiver effective noise temperature, K

Evaluating the constants for Eq. (1)

$$B = \frac{T_4 - T_1}{R_4 - R_1} \tag{2}$$

$$A = \frac{T_1 R_4 - T_4 R_1}{R_4 - R_1} \tag{3}$$

or alternatively

$$A = T_1 - BR_1 \tag{4}$$

where

T_1, T_4 = system input calibration parameters

R_1, R_4 = system output readings with inputs T_1 and T_4

For a microwave radiometer system, the solution for B can be accomplished by switching the receiver input between two matched terminations at known physical temperatures T_{p4} and T_{p1}, and readings R_4 and R_1. The usual assumption is that $R = 0$ when $T = 0$. For this case

$$B = \frac{T_{p4} - T_{p1}}{R_4 - R_1} \tag{5}$$

$$A = 0 \tag{6}$$

A typical DSN microwave radiometer system is presently configured with a single microwave ambient termination at physical temperature T_{p4} so that Eq. (5) is not applicable. Also, Eq. (6) may not be exactly applicable if $R_1 \neq 0$ when $T_1 = 0$. It is possible to simulate T_1 by terminating the reading device input with an ambient termination (at temperature T_{p1}) and either setting $R_1 = 0$ (zero set) or reading R_1 (bias). In this case, T_1 is estimated using $T_1 = T_{p1}/G$, where G is the overall gain between the microwave system input and the reading device input termination. Assuming that T_e (defined from receiver input to the reading device input termination) is known from a laboratory measurement, we have, for this case

$$B = \frac{T_{p4} + T_e - T_1}{R_4 - R_1} \tag{7}$$

$$A = T_1 - BR_1 \tag{8}$$

$$T_1 = \frac{T_{p1}}{G} << 1 \tag{9}$$

assuming G is much greater than 1.

For $T_1 = R_1 = 0$

$$B = \frac{T_{p4} + T_e}{R_4} \tag{10}$$

$$A = 0 \tag{11}$$

Eq. (1) is now used to obtain the system temperature on the antenna T_2, from the system output reading R_2, using A and B as calculated above.

$$T_{op} = T_2 = A + BR_2$$

B. Non-Linear Analysis

The linear analysis, Eq. (1), will be in error if the system is non-linear. For small non-linearities, assume a quadratic solution (see Fig. 1). Other models are also possible; a quadratic was chosen for this study since it is the next term in a power series, it is consistent with previous studies [3], and good results are obtained with existing microwave radiometer systems.

$$T = A + BR + CR^2 \tag{12}$$

There are various practical techniques for measuring A, B, and C. For a non-linear system, constants A and B from Eq. (12) are not the same as in Eq. (1). Consider the example of a

microwave radiometer system with a manually switched noise diode which adds a temperature T_n to the microwave system input noise temperature when turned on. There are five conditions to consider, with a noise temperature and reading for each. Assume that readings R are taken for each case. Then for the five conditions we find the following input values, T, and output readings, R (see Fig. 1):

(1) calibrated termination, ND off: $T = T_1; R = R_1$

(2) antenna, ND off: $T = T_2; R = R_2$

(3) antenna, ND on: $T = T_3 = T_2 + T_n; R = R_3$

(4) calibrated termination, ND off: $T = T_4; R = R_4$

(5) calibrated termination, ND on: $T = T_5 = T_4 + T_n$; $R = R_5$

Using these conditions with Eq. (12)

$$A + BR_1 + CR_1^2 = T_1 \tag{13}$$

$$A + BR_2 + CR_2^2 = T_2 \tag{14}$$

$$A + BR_3 + CR_3^2 = T_2 + T_n \tag{15}$$

$$A + BR_4 + CR_4^2 = T_4 \tag{16}$$

$$A + BR_5 + CR_5^2 = T_4 + T_n \tag{17}$$

Solving five equations with knowns T_1, T_4, R_1, R_2, R_3, R_4, and R_5 and 5 unknowns, A, B, C, T_2 and T_n:

$$C = \frac{T_4 - T_1}{R_4^2 - R_1^2 - (R_4 - R_1)D} \tag{18}$$

$$B = -CD \tag{19}$$

$$A = T_1 - BR_1 - CR_1^2 \tag{20}$$

$$T_2 = A + BR_2 + CR_2^2 \tag{21}$$

$$T_n = A + BR_3 + CR_3^2 - T_2 \tag{22}$$

where

$$D = \frac{R_5^2 - R_4^2 - R_3^2 + R_2^2}{R_5 - R_4 - R_3 + R_2}$$

T_4 and T_1 are determined for the appropriate configuration, such as a microwave radiometer system, using the same concepts illustrated with Eq. (5).

The A, B, and C constants adjust the non-linear curve to fit the equal input temperature level changes, $T_n = (T_5 - T_4) = (T_3 - T_2)$ for the observed system output readings. This is the central concept of the proposed technique. With a curve defined by the solution of A, B, and C, the "primary" unknown, T_2, is solved with Eq. (21).

These results are general, and can be used either to provide an estimate of system non-linearity or to apply a correction. It is useful to determine the difference between T_2 provided by the linear and non-linear analyses, Eqs. (1) and (12). An assumption that the non-linear analysis corrects for non-linearity and the linear analysis does not can be used to obtain an estimate of the non-linearity error. Defining a linearity factor:

$$\begin{aligned} LF &= T \text{ (linear analysis)}/T \text{ (non-linear analysis)} \\ &= T \text{ (computed from Eq. [1])}/T \text{ (computed from Eq. [12])} \end{aligned} \tag{23}$$

Typically, LF is greater than 1 for an NAR system and less than 1 for a total power system. The results section has further discussion of the linearity factor.

It is also useful to consider a linearity correction factor (CF) for difference measurements based on Eq. (12). This is useful for radio source noise temperature measurements. Assume that the source measurement corrected with Eq. (12) is given by

$$T_{sc} = T_{su} CF \tag{24}$$

where

$CF = B + 2CR_{off} + CT_{su}$ (CF is a ratio obtained from Eq. [12])

$T_{su} = R_{on} - R_{off}$ = uncorrected source measurement, K

R_{off} = uncorrected off source measurement

R_{on} = uncorrected on source measurement

The best strategy for correction of source noise temperature difference measurements depends on the requirements and instrumentation. In some cases the correction should be made in near real time, during the sequence of measurements. In other cases, particularly with small corrections, the correction could be made after the sequence of measurements. Consider an example with $B = 1.02$, $C = 0.0001$, $T_{su} = 1$ K, and R_{off} values of 20 K to 40 K with an elevation angle over the

measurement sequence. The corrected source noise temperature difference value computed with the averaged $R_{off} = 30$ K is $T_{sc} = 1.026$ K. Individual corrected values would range from 1.024 K to 1.028 K. The peak "error" in the single correction relative to the full range of uncorrected measurements is only about 0.01 dB for this example. This indicates that a simple single correction may be satisfactory for many applications.

III. Results

The above equations have been programmed for an IBM compatible personal computer (PC) SuperCalc computer program, assuming a single temperature-calibrated ambient microwave termination and a known receiver noise temperature, T_e. Noise temperature measurements were made to verify the procedures using the Goldstone DSS 13 antenna at 2.3 GHz (S-band). A typical data set taken July, 1987, is shown in Fig. 2. Data sets 101, the NAR case, and 102, the total power case, are time meshed together in order to minimize other than nonlinearity effects. For this data, $T_1 \approx 0$, since there is high amplifier gain between the antenna and the square law detector. The computer program allows finite values for T_1 and R_1 ("bias") if desired.

The NAR case used an IBM compatible PC configuration. An S-band HEMT low noise amplifier was followed by a mixer, an IF amplifier, a square law diode detector assembly, and an analog voltage to digital interface to the PC. The PC controlled the periodic on-off switching of the NAR noise diode at the HEMT amplifier input. The noise diode required for non-linearity calibration was turned on and off manually as required. The total power case used a voltmeter connected directly to the analog voltage output of the square law detector.

Eqs. (1) and (12) are used for the linear and non-linear solutions tabulated in Table 1. One-sigma statistical measurement errors are in parentheses. The A, B, and C constants are quite different for the linear and non-linear analyses. This indicates significant non-linearity for these configurations. For the non-linear analysis system temperature on the antenna T_2, results for both the NAR and total power cases are in good agreement.

The linear and non-linear analysis difference for the NAR case, data set 101, is shown plotted in Fig. 3 as the non-linearity error. The error goes to zero on the ambient load since this is a common reference for both the linear and non-linear analysis. The linearity factor, LF (Eq. [23]), is shown plotted in Fig. 4. A ratio of 1.03, obtained on the antenna, indicates about 3 percent non-linearity. This linearity factor is included in Table 1 for both the NAR and total power cases, indicating an approximate equal magnitude for each. Figures 3 and 4 include the plus and minus 1-sigma statistical measurement error obtained from the 6 independent data measurements. Most of the non-linearity of this configuration is due to the square law diode detector. This fact was determined by reducing the power level to the detector and repeating the data set and analysis. Reducing the power level to the square law diode reduces the non-linearity error at the expense of increased zero drift in the dc amplifiers following the square law diode, and less resolution in the analog to digital converters.

Figure 5 shows a plot of the differences between the linear analysis, Eq. (1), and non-linear analysis, Eq. (12), vs. T_{op} for both the NAR case, data 101, and total power case, data 102. The small differences between the two different cases after correction for non-linearity indicate successful removal of the non-linear effects between two separate radiometer operating modes.

The zenith system noise temperature of the DSS 13 S-band receiving system with a HEMT low noise amplifier was remeasured by station personnel on August 28, 1987, with three different time-interleaved data types. The data cases were: (1) IBM PC NAR with a square law detector and A/D converter; (2) total power with a square law detector and digital voltmeter; and (3) total power with a digital power meter. The results are tabulated in Table 2 and plotted in Fig. 6. This shows poor agreement between the data types using the linear analysis, Eq. (1). The non-linear analysis, Eq. (12), corrects for system non-linearity and provides good agreement between the three data types, increasing confidence in the technique. It is assumed that the difference between the corrected values for the July 2 and August 28, 1987, data is partly due to different atmospheric conditions and partly to other measurement bias errors.

IV. Conclusion

A technique for the determination and removal of system non-linearity has been analyzed and demonstrated. This includes the concept, equations, and results applicable for DSN microwave noise temperature measurements. Comparison corrections for separate radiometer modes of operation resulted in very nearly equal system temperature after correction for the system non-linearity. The primary source of radiometer system non-linearity was shown to be in the square law detector. The best remedy for system non-linearity is proper equipment design and system power level settings. Unfortunately, this is not always achieved. With a given implemented system and operating levels, the above techniques provide a useful tool for estimating the system instrumentation non-linearity and correcting for the non-linearity if necessary. A correction may not be advisable if it is of the same magnitude as the correction measurement statistical error, or if it is small when compared to other error sources.

Acknowledgment

R. Riggs provided stimulating discussions and unpublished computer simulations for the receiver and square law detector model. Many data sets were taken by DSS 13 personnel (J. Garnica, *et al.*) as well as by some DSS 15 personnel (R. Caswell, *et al.*). L. Skjerve provided the IBM compatible PC NAR configuration used for the DSS 13 data.

References

[1] P. D. Batelaan, R. M. Goldstein, and C. T. Stelzried, "A Noise Adding Radiometer for Use in the DSN," *JPL Space Programs Summary 37-65*, vol. 2, Jet Propulsion Laboratory, Pasadena, California, pp. 66–69, September 30, 1970.

[2] C. T. Stelzried, *The Deep Space Network–Noise Temperature Concepts, Measurements, and Performance*, JPL Publication 82-33, Jet Propulsion Laboratory, Pasadena, California, September 15, 1982.

[3] A. J. Freiley, J. E. Ohlson, and B. L. Seidel, "Absolute Flux Density Calibrations: Receiver Saturation Effects," *DSN Progress Report 42-46*, vol. May–June 1978, Jet Propulsion Laboratory, Pasadena, California, pp. 123–129, August 15, 1978.

[4] C. T. Stelzried, "Noise Adding Radiometer Performance Analysis," *TDA Progress Report 42-59*, vol. July–August 1980, Jet Propulsion Laboratory, Pasadena, California, pp. 98–106, October 15, 1980.

[5] "IRE Standards on Electron Tubes: Definitions of Terms, 1962 (62 IRE 7.S2)," *Proc. IEEE*, pp. 434–442, March 1963.

Table 1. DSS 13, July 2, 1987, S-band, zenith, noise temperature measurement analysis results (1-sigma accuracy shown in parentheses)

Radiometer case	Parameter	Analysis method	
		Linear, Eq. (1)	Non-linear, Eq. (12)
	A	0	0
	B	1.0034 (0.0004)	1.0413 (0.0010)
	C	–	−0.00012 (0.000002)
NAR (data set 101)	Noise diode, T_n, K (on antenna)	56.4 (0.02)	57.7 (0.04)
	Noise diode, T_n, K (on ambient load)	60.4 (0.09)	57.7 (0.04)
	System noise temperature on antenna, T_2, K	31.3 (0.02)	32.3 (0.03)
	Linearity factor on antenna, LF (ratio)		1.034 (0.001)
	A	0	0
	B	281.23 (0.54)	268.36 (0.59)
	C	–	11.82 (0.20)
Total power with digital voltmeter (data set 102)	Noise diode, T_n, K (on antenna)	59.3 (0.07)	57.7 (0.05)
	Noise diode, T_n, K (on ambient load)	54.8 (0.04)	57.7 (0.05)
	System noise temperature on antenna, T_2, K	33.7 (0.04)	32.4 (0.05)
	Linearity factor on antenna, LF (ratio)		0.959 (0.001)

Table 2. DSS 13, August 8, 1987, S-band, zenith, noise temperature measurement analysis results (1-sigma accuracy shown in parentheses)

Radiometer case	Parameter	Analysis method	
		Linear, Eq. (1)	Non-linear, Eq. (12)
NAR (data set 101)	System noise temperature on antenna, T_2, K	30.8 (0.02)	31.9 (0.10)
	Linearity factor on antenna, LF (ratio)		1.034 (0.003)
Total power with digital voltmeter (data set 102)	System noise temperature on antenna, T_2, K	32.9 (0.13)	31.9 (0.19)
	Linearity factor on antenna, LF (ratio)		0.969 (0.003)
Total power with power meter	System noise temperature on antenna, T_2, K	31.9 (0.04)	31.8 (0.10)
	Linearity factor on antenna, LF (ratio)		0.997 (0.002)

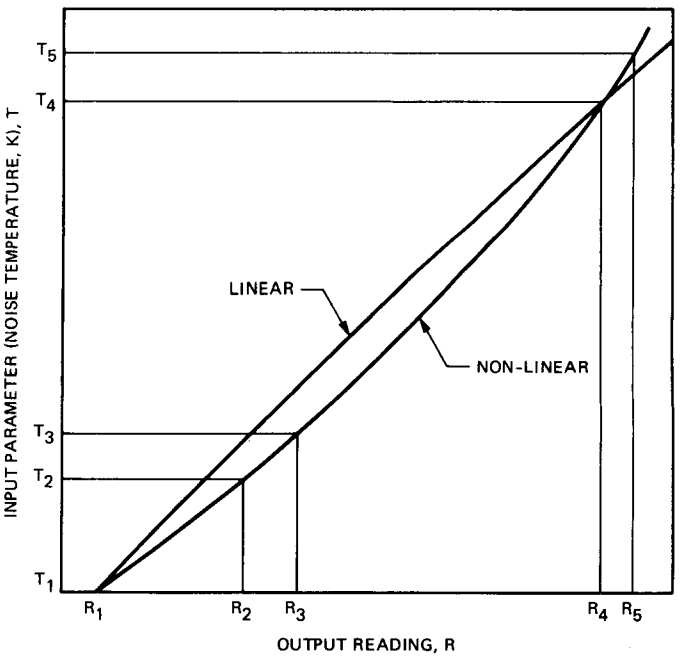


Fig. 1. Representative diagram of non-linear system input parameter vs. output reading, assuming that $T_n = (T_5 - T_4) = (T_3 - T_2)$

LINEARITY CALIBRATION DATA SHEET (CTS LIN2, 4-30-87)

OPERATOR C. GOODSON FREQUENCY, MHz 2295 DATE 7/2/87

CONFIGURATION 1.8 V DETECTOR: SN 536

T_f, K 0.471 T_m (MASER OR LNA), K 10 INTEGRATION TIME, SEC 40

T_N (IF USING NAR), K 76.6 WEATHER CLEAR

	0.00	0.00	0.00	0.00	0.00	0.00
1.	0.00	0.00	0.00	0.00	0.00	0.00
R_1	(DETECTOR OR POWER METER IN Z_0)					
	0.121	0.120	0.120	0.120	0.120	0.119
2.	31.136	31.108	31.119	31.173	31.193	31.281
R_2	(ANTENNA)					
	0.333	0.332	0.332	0.330	0.330	0.329
3.	87.245	87.295	87.275	87.353	87.391	87.478
R_3	(ANTENNA AND NOISE DIODE)					
	1.095	1.093	1.089	1.089	1.084	1.083
4.	304.887	304.858	304.576	305.718	305.325	305.551
R_4	(AMBIENT TERMINATION)					
	1.291	1.288	1.284	1.284	1.278	1.277
5.	365.041	365.110	364.961	365.632	365.374	365.916
R_5	(AMBIENT TERMINATION AND NOISE DIODE)					
6.	22.23	22.40	22.50	22.64	22.77	22.88
R_6	(T_p, C)					
7.	1734	1740	1745	1753	1759	1806
(TIME, UT)						

→ TOTAL POWER CASE, DATA 102 (TYP)

→ IBM NAR CASE, DATA 101 (TYP)

Fig. 2. Typical radiometer noise temperature data set (time meshed NAR and total power cases) taken July 2, 1987, at DSS 13, with 2.3-GHz HEMT low noise amplifier

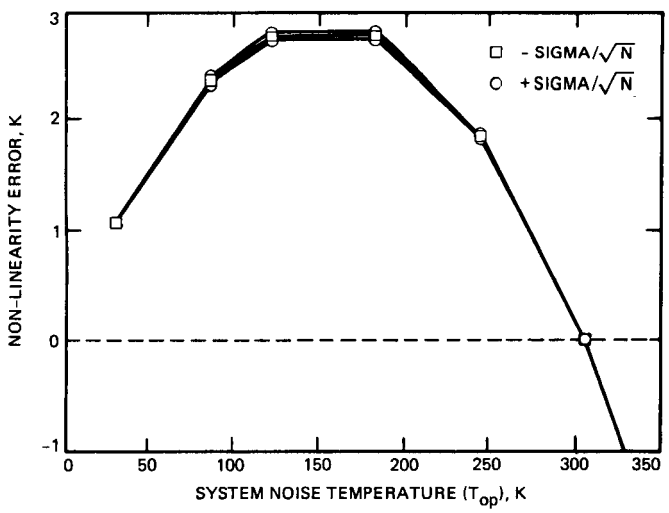


Fig. 3. Plot of the difference noise temperature for the NAR case, data set 101, between the non-linear and linear analyses as a function of system noise temperature

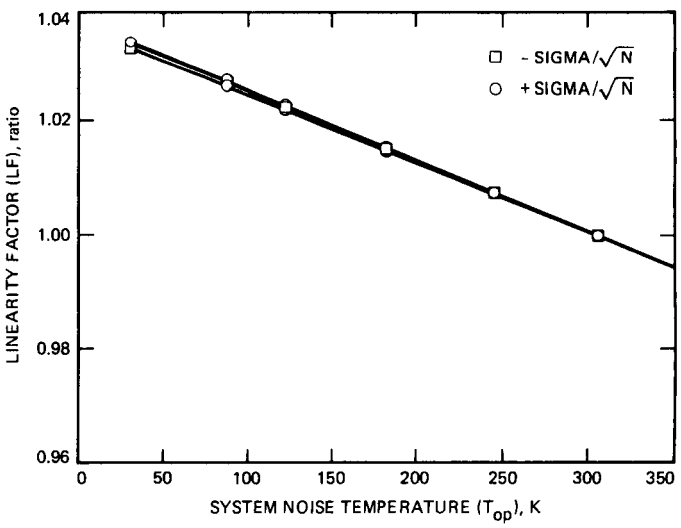


Fig. 4. Plot of the system linearity factor, LF (Eq. [23]), for the NAR case, data set 101, as a function of system noise temperature

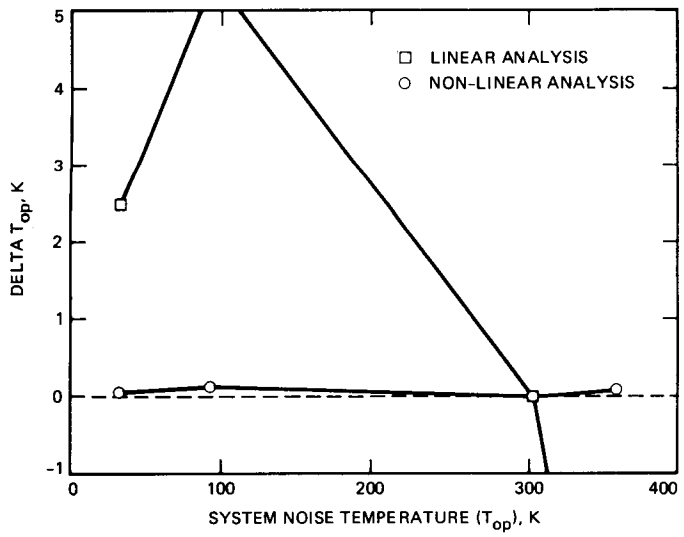


Fig. 5. Plot of the differences, ΔT_{op}, between the NAR and total power cases for both linear and non-linear analyses as a function of system noise temperature (see Table 1)

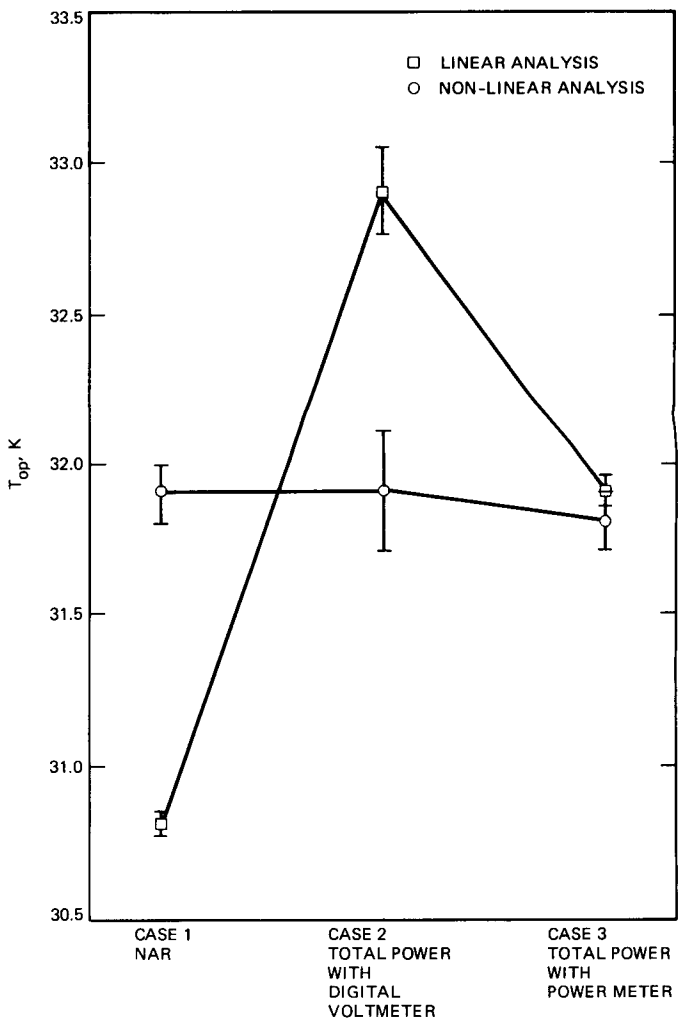


Fig. 6. Plot of the August 28, 1987, zenith, 2.3-GHz noise temperature measurements using the NAR and total power cases showing linear and non-linear analysis results (see Table 2)

Atomic Frequency Standards for Ultra-High-Frequency Stability

L. Maleki, J. D. Prestage, and G. J. Dick
Communications Systems Research Section

In this article, the general features of the $^{199}Hg^{+}$ trapped-ion frequency standard are outlined and compared to other atomic frequency standards, especially the hydrogen maser. The points discussed are those which make the trapped $^{199}Hg^{+}$ standard attractive–high line Q, reduced sensitivity to external magnetic fields, and simplicity of state selection, among others.

I. Introduction

Since the inception of the DSN, the need for stable frequencies to support navigation and certain radio science experiments has been well recognized. In the early days of the DSN, when missions were limited to lunar distances, the precision required for navigational parameters implied that stabilities on the order of one part in 10^{10} would be adequate for all applications. For this level of performance, the quartz crystal oscillator was quite adequate.

The more stringent requirements of missions exploring the planets, however, modified this perception. The advent of VLBI further accentuated the need for more stable frequencies, and thus the application of atomic frequency standards became necessary at the DSN.

The first such standards were the rubidium cell devices, with stabilities in the range of 5×10^{-13} over averaging intervals of 10^2 to 10^4 seconds. Stabilities provided by these standards were inadequate for a number of applications that subsequently emerged, and thus these standards soon gave way to the more stable hydrogen masers. Hydrogen masers then became the primary frequency standards throughout the DSN, providing users with stable reference signals for navigation and radio sciences.

A number of problems with hydrogen masers have been encountered since the early years of their application in the DSN. The major problem associated with masers relates to reliability. Hydrogen masers experience unexpected failures that are usually associated with the large flow of hydrogen gas into the vacuum pumps. Despite ongoing attempts to predict or prevent these failures, at present the only certain way to assure the DSN of the availability of reference signals is through redundancy. This step is, however, expensive, since hydrogen masers currently cost approximately $450K.

Aside from these reliability issues, a number of more stringent requirements for frequency stability have emerged in the past few years which exceed the current capability of hydrogen masers. The most notable of these requirements is the need for stabilities in the range of one part in 10^{17}—a level nearly two orders of magnitude beyond the stability of masers currently employed in the DSN.

The cost associated with the procurement and maintenance of the masers, together with the need for better frequency stabilities, led to an effort to identify technologies which could provide the DSN with more attractive alternatives. It was soon recognized that confined-ion technology had the potential to provide a viable alternative to hydrogen masers. This led to the establishment of a task, under the Advanced Systems Program of the TDA, to carry out a research-and-development effort aimed at the production of a mercury standard based on ion confinement. This effort recently culminated in the demonstration of a mercury ion frequency standard, described in detail in a separate article [1]. The purpose of the present article is to outline the technical background of atomic frequency standards and to identify the features which render the trapped-ion device attractive for frequency-standard applications.

The first part of this article will describe the general characteristics of atomic frequency standards. The major portion of this discussion will center on a description of the hydrogen maser, with which there is a more general familiarity. The description of trapped-ion devices will be presented in parallel with that of the maser in order to elucidate features which characterize each device. The final segment of the article will focus on specific characteristics of trapped-ion devices.

II. General Characteristics of Atomic Standards

Atomic frequency standards are based on the discrete energy structure of atomic systems (atoms and ions). A pair of discrete energy levels may be connected by a photon, the energy of which is equal to the energy separation of the two levels. This may be represented as $(E_u - E_l) = hf$, where E_u and E_l are the energies of the upper and lower levels, respectively; h is Planck's constant; and f is the frequency of the photon.

For an excited state, there is a finite probability that the atom may remain in that state before decaying to a state of lower energy by emission of a photon. The manifestation of this probability is the natural energy width associated with every excited state. The frequency of the emitted photons then exhibits a finite width corresponding to the energy width of the upper state.

This quantum mechanical condition provides a recipe for establishing precise sources of frequencies based on atomic structures. One selects an atomic system containing a suitable pair of energy levels, with the upper level having a narrow width. The frequency of the photon connecting the two levels will then define the precise frequency on which the standard could be based.

The fundamental limitation on atomic frequency standard operation is imposed by its operating frequency, ν_0; by the number of atoms or ions available, N; and by the interrogation time, T. Since the radiation-limited lifetime of atomic-microwave transitions is essentially infinite (10^4 years is typical), an effective limit is imposed instead by the interrogation time in which the frequency source is required to operate. For measurements shorter than (and equal to) this interrogation time, a secondary frequency source must be provided. This limit can be expressed as

$$\sigma_y(\tau) = \frac{1}{2\nu_0}\sqrt{\frac{1}{NT}}\sqrt{\frac{1}{\tau}}$$

From this it is seen that high stability is achieved by the following factors: high frequency operation allows a small fractional linewidth to be determined within the allowed interrogation time; many particles yield good statistics to determine the frequency even within that linewidth; and well isolated atoms (together with an excellent local oscillator) allow a long interrogation time.

The simplicity of the principle outlined above greatly distorts the degree of difficulty encountered in actual implementation. For one thing, statements made in the paragraphs above pertain to "unperturbed" atoms. This signifies atoms that are free of the influence of all external fields and free of interaction with all external systems. Such an ideal situation is difficult to realize, and thus one would have to be content with providing an environment which minimizes the influence of external fields and interactions. Furthermore, the frequency of the photon should be in a practical range that is easily accessible for measurement and conversion. At the current level of technology, this implies frequencies in the microwave regime, limiting the suitable energy states to the hyperfine levels of atomic systems, or vibrational levels of molecules.

These limitations reduce the number of candidate atomic systems suitable for frequency-standard application to the hyperfine levels of alkali and alkali-like atoms. Alkali atoms, such as rubidium and cesium, have a single electron outside a closed shell, resulting in a simple ground-state hyperfine structure. This characteristic is also a feature of the hydrogen atom and the mercury ion.

Atomic systems are utilized in two modes for frequency-standard applications. In the active mode, the atom acts as an oscillator, emitting radiation with the characteristic sharp center frequency and narrow linewidth that is then used to stabilize the frequency of a slaved oscillator. In the passive mode, the role of the atom is much like that of a filter for the absorbed radiation derived from the slaved oscillator. Exam-

ples of active atomic frequency standards include the hydrogen maser and the rubidium maser. Passive standards include the cesium beam clock, the rubidium gas cell standard, and the trapped mercury ion standard. The variance of the stability for active sources characteristically shows a $1/\tau$ dependence on the measuring time, while that for passive sources shows a $1/\sqrt{\tau}$ dependence. The $1/\tau$ behavior results from follower amplifier noise added to the very weak microwave signal available from the atomic radiators. In practical terms, an active source (the hydrogen maser) shows rapidly increasing stability at moderate averaging times ($\tau < 1000$ seconds), while passive sources show improvement that is slower but that extends to longer averaging times ($\tau > 10{,}000$ seconds).

III. Operational Characteristics

Whether the atomic system is applied in the active or passive mode, it is first necessary to prepare the atom in the appropriate hyperfine state. In thermal equilibrium, the population of energy levels of the atom is governed by the Boltzmann distribution for a given temperature. At room temperature, hyperfine levels of the ground state have nearly equal populations owing to the relatively small differences in their corresponding energy splitting compared to the thermal energies. In order to absorb or emit photons at the hyperfine frequency f, a population difference between the levels must be created.

The atoms in the desired hyperfine state may be selected either by rejecting the ones in the undesired state or by converting all atoms to the desired state. In the case of the hydrogen maser and the cesium clock, this preparation is accomplished by passing the beam of atoms through an inhomogeneous magnetic field which focuses the desired species and rejects atoms in the unwanted states by diverting them out of the beam. This action is due to the interaction of the atomic magnetic moments, corresponding to a given hyperfine state, with the magnetic field gradient.

For the rubidium cell standard and the trapped mercury ion device, atoms in the unwanted state are converted to the desired state via a process called optical pumping. In this process, atoms in the undesired hyperfine state absorb an optical photon from a resonant light source and are promoted to an excited energy state. The excited state is selected such that the excitation is short-lived (typically lasting a few nanoseconds) and transitions are allowed to both hyperfine states. Since the excitation takes place out of the undesired state but the decay of the optically excited states is to both hyperfine levels, the population of the undesired state is depleted, and eventually all atoms absorbing the pump light end up in the desired hyperfine state.

Optical pumping of rubidium and cesium is made possible by dye lasers and semiconductor lasers. Use of lasers for optical pumping of mercury is not practical because the required UV radiation has a wavelength of 194 nm, which is outside the range of present-day lasers. Frequency mixing techniques are possible, however, and the time and frequency research group within the National Bureau of Standards has demonstrated laser pumping of mercury ions [2].

After state selection, the energy structure of the atomic system is interrogated for the precise frequency. In the case of the hydrogen maser, the beam of atoms in the upper hyperfine state enters a high-Q cavity which is tuned to resonate at the hydrogen atom hyperfine energy difference (1.4 GHz). The oscillation in this cavity coherently stimulates the emission of more 1.4-GHz radiation from the atoms into the cavity, overcoming its losses and sustaining the maser oscillation.

The atomic linewidth is limited by the interrogation time, which for the maser corresponds to the time the atoms are contained in the high-Q cavity. The atoms are contained for about 1 second inside a Teflon-coated quartz bulb, giving 1 Hz atomic linewidth. The bulb occupies a region of the cavity where the microwave field has nearly constant phase and is smaller than the wavelength of the emitted radiation, thus eliminating the first order Doppler effect through a process called Dicke narrowing [3].

During the 1 second storage time, the hydrogen atoms undergo many collisions with the walls of the quartz bulb; they also undergo spin exchange collisions with each other as well as with other atomic and molecular species that may be present in the vacuum system. These collisions limit the performance of the maser in a number of ways. The hydrogen atom energy levels are shifted because of the large local electromagnetic fields encountered during collisions. This generates a frequency offset from the ideal unperturbed hydrogen atom that will vary from bulb to bulb and with the density of hydrogen atoms in the bulb.

Collisions with the walls of the quartz bulb also constitute a loss mechanism for the atoms in the desired hyperfine state; recombination and other chemical reactions reduce the number of excited atoms. One consequence of this loss is a reduction in the power level output of the maser and an associated reduction in the signal-to-noise ratio. Ions contained in an electromagnetic trap are free of perturbations associated with wall and spin exchange collisions. Ions generated by the impact of electrons on a background gas are trapped in a small region of space, typically smaller than a cubic centimeter. The confinement is produced through the application of appropriate electromagnetic fields to a three-electrode trap structure with the appropriate geometry. The

fields may be static electric and static magnetic (as employed in Penning traps), or they may be rf electric fields (employed in the Paul's or rf trap). Depending on the geometry of the trap and the value of the field parameters, a particle with a given ratio of charge/mass can execute closed orbits and be trapped. Like the quartz bulb in the hydrogen maser, the ion trap confines the ions to a region smaller than the wavelength of the 40.5 GHz radiation (~7.5 mm), thus eliminating the first order Doppler effect.

The problem of wall collisions and their associated frequency shift is thus eliminated for trapped ions. Furthermore, the small density of the ions (on the order of 10^6 per cubic centimeter) resulting from the space charge effect virtually eliminates de-phasing collisions with other confined ions. Finally, the confinement time of the ions is governed by the density of the background gas, which may increase the kinetic energy of the ions through collisions and knock them out of the trap. The confinement time, however, can be made as long as many hours, depending on the vacuum conditions. This means that narrower lines corresponding to longer interrogation times may be obtained with trapped ions.

Since one of the parameters governing the stability of the atomic frequency standard is the line Q (where $Q = \Delta f/f_0$, with Δf the linewidth and f_0 the center frequency), a transition with a larger center frequency yields a higher attainable stability for the same Δf and the same signal-to-noise ratio. Thus the line Q of the hydrogen maser is typically a few times 10^9, corresponding to the linewidth of a hertz and a transition frequency of 1.4 GHz. The hyperfine transition of mercury ions (isotope 199), by comparison, is about 40.5 GHz. A linewidth of one hertz produces a Q which is nearly a factor of 30 larger than the corresponding Q of hydrogen. This feature, together with the simple hyperfine structure of the ground state and the possibility of optical pumping with a discharge lamp, renders mercury ions particularly suitable for trapped-ion standards.

Two other characteristics of mercury supplement the features mentioned above. The large mass of the mercury ion implies a small frequency shift due to the second order Doppler effect. This motional effect is due to relativistic time dilation, whereby the frequency of a moving ion is slightly shifted with respect to an observer in the laboratory frame of reference. For particles with a given temperature and thermal energy, those with the largest mass have the smallest speed and thus the smallest second order Doppler shift. This effect represents the ultimate limitation on the long term frequency stability of the trapped mercury ion standard. The size of the effect at room temperature is on the order of one part in 10^{13} for room temperature Hg^+ ions. However, because the trapping force in an rf trap is generated by the *motion* of charged particles in an inhomogeneous oscillating electric field, 10^6 Hg^+ ions held under typical trapping conditions will have a second order Doppler shift of about one part in 10^{12} [4]. Trap designs which do not increase the second order Doppler shift in this way but that have 10^6 or more ions will be discussed in a future article. Obviously, cooling the particles will lead to the reduction of the second order Doppler effect. Here again, trapped ions lend themselves to drag cooling produced through collisions with a light, inert gas such as helium or by laser cooling. The latter process amounts to the loss of kinetic energy through momentum transfer associated with the directional absorption of photons, followed by the emission of photons in all directions. Such an approach requires a laser at 194 nm, and the NBS group [2] has demonstrated laser-cooled mercury.

Finally, the ease of introduction of mercury in the vacuum system is an important consideration. Isotopic mercury is easily liberated from oxide material with negligible vapor in the needed quantities. Too much background vapor of mercury reduces the trap time through collisions, while too little requires longer periods for loading the trap with ions. Operating pressures of 10^{-9} torr are readily achieved in the trap. This reduces the gas load to the vacuum system, allowing long-term and reliable operation with small sorption or ion pumps.

By contrast, the production of hydrogen atoms from naturally occurring hydrogen in the molecular form requires the use of an rf-powered dissociator. The dissociator has an efficiency of atom production typically in the five percent range. The aging of the dissociator also represented a source of failure for the early masers. The low efficiency of atom production and the required high flux of the continuously flowing hydrogen produce a significant load for the maser vacuum pumps, which are required to maintain a background pressure in the 10^{-6} torr range. The reliable operation of these pumps, also of the chemical sorption or ion variety, represents the major challenge associated with hydrogen masers.

IV. Sensitivity to Environmental Perturbations

Atomic frequency standards are generally extremely sensitive to environmental influences. Ambient electromagnetic fields perturb the structure of the energy levels and the corresponding frequencies of transition. Thus, the presence of magnetic fields disturbs the atomic level via the Zeeman effect. The Stark effect reflects the interaction of electric fields with energy states. Strong electromagnetic radiation interacts with the atomic levels and results in the "light shift," or the dynamic Stark effect.

The degree of perturbation produced by the ambient fields naturally depends on the intensity of the fields. However, the interaction is also dependent on the particular level of the particular atomic system. Thus, for example, the dependence of the frequency shift of a trapped mercury ion on ambient magnetic field variations is some 30 times smaller than the corresponding shift in the maser.

Another important source of frequency drift for the maser is the so-called cavity pulling effect. This effect is based on the fact that the maser frequency is related to the hydrogen-atom frequency and the cavity frequency. Any temperature drifts that can cause dimensional changes in the cavity will change the cavity resonance frequency, which in turn will alter the maser frequency. This effect may be reduced through control of the temperature of the cavity, which is chosen from materials already low in their coefficients of thermal expansion.

Nevertheless, the sensitivity of the maser frequency is high enough to detect shifts corresponding to dimensional changes of the cavity at the 0.25 angstrom level. The problem is stabilized to some extent through various schemes for "autotuning" of the cavity [5]. Other environmental effects, such as changes in barometric pressure and even in humidity, have been observed to influence sensitive components to produce frequency drifts [6]. The same mechanism for frequency drift due to dimensional changes does not exist for the trapped mercury ion standard, since there is no resonant cavity. An important ramification of this fact is that bulky temperature controlling shields and ovens are not required. This feature simplifies the structure and reduces the associated mass of the device.

V. Trapped-Ion Technical Aspects

In the preceding sections, general features of atomic frequency standards were considered, and comparisons were made between the hydrogen maser and the trapped mercury ion device. It was indicated that because of its inherent properties, the trapped-ion device offers the potential for a frequency standard which is simpler in structure, has less mass and size, is more reliable, and has better stability performance.

Nevertheless, the trapped mercury ion standard has a few features which pose technological challenges that must be met before the full potential of this device can be realized. Some of the features unique to the trapped mercury ion are described in this section.

The confinement time of ions depends on a number of parameters, the most significant of which is the collision rate of mercury ions with background mercury vapor and with residual impurities in the vacuum system. These collisions may also redistribute the population of levels prepared by optical pumping. Consequently, a background pressure of less than 10^{-9} torr is desired to diminish the influence of background particles. This level of vacuum is not difficult to realize but requires care appropriate to ultra-high vacuum practices.

Another unique requirement for the trapped mercury ion standard relates to the lamp used for optical pumping. While much has become known about mercury lamps through the experience of the lighting industry over the years, the requirements of the light source for optical pumping are rather different. Here the radiation of interest is due to ion emission, while the mercury neutral emission copiously produced and used in fluorescent lights is the unwanted background. Light emission from the ions implies a higher plasma temperature in the lamp, which in turn requires higher input power. Thus, about 20 watts of rf power is coupled inductively to an rf excited lamp, resulting in accelerated damage to the glass and ultimately in short operating life. A number of design approaches, potential solutions, and candidate technologies are being used to address the problem of lamp life. Nevertheless, the development of efficient and reliable lamps continues to challenge groups active in this work. The frequency stability of the trapped-ion device depends on the signal-to-noise ratio of the scattered 194 nm radiation. An increase in this parameter directly improves the ultimate stability. Improvement of the input optics and collection optics, together with all other steps that enhance the signal-to-noise ratio, will continue to be pursued by workers in this field. A number of approaches leading to increased efficiency of microwave photon detection through the detection of the UV radiation at 194 nm are currently under investigation.

Finally, mention was made of the sources of drift in the device, particularly the offset due to the second order Doppler effect. Cooling of the ions is an approach which reduces most unwanted drift. At the present time, however, a complete and full understanding of all sources of ion heating, both collisional and that due to rf fields, is lacking.

References

[1] J. D. Prestage, G. J. Dick, and L. Maleki, "The JPL Trapped Mercury Ion Frequency Standard," *TDA Progress Report 42-91*, vol. July–September 1987, Jet Propulsion Laboratory, Pasadena, California, November 15, 1987.

[2] J. C. Bergquist, W. M. Itano, and D. J. Wineland, "Recoilless Optical Absorption and Doppler Sidebands of a Single Trapped Ion," *Phys. Rev.*, vol. A-36, pp. 428–430, July 1987.

[3] R. H. Dicke, "The Effect of Collisions upon the Doppler Width of Spectral Lines," *Phys. Rev.*, vol. 89, pp. 472–473, January 1953.

[4] L. S. Cutler, R. P. Giffard, and M. D. McGuire, "Thermalization of ^{199}Hg Ion Macromotion by a Light Background Gas in an RF Quadrupole Trap," *Appl. Phys.*, vol. B-36, pp. 137–142, 1985.

[5] G. J. Dick and T. K. Tucker, "Fast Autotuning of a Hydrogen Maser by Cavity Q Modulations," *TDA Progress Report 42-91*, vol. July–September 1987, Jet Propulsion Laboratory, Pasadena, California, November 15, 1987.

Measurement and Analysis of Cryogenic Sapphire Dielectric Resonators and DROs

G. J. Dick
Communications Systems Research Section

This article presents the experimental and computational results of a study on a new kind of dielectric resonator oscillator (DRO). It consists of a cooled, cylindrically symmetric sapphire resonator surrounded by a metallic shield and is capable of higher Q's than any other dielectric resonator. Isolation of fields to the sapphire by the special nature of the electromagnetic mode allows the very low loss of the sapphire itself to be expressed. Calculations show that the plethora of modes in such resonators can be effectively reduced through the use of a ring resonator with appropriate dimensions. Experimental results show Q's ranging from 3×10^8 at 77 K to 10^9 at 4.2 K. Performance is estimated for several types of DROs incorporating these resonators. Phase noise reductions in X-band sources are indicated at values substantially lower than those previously available.

I. Introduction

A new kind of dielectric resonator promises to enable an important advance in the capability of dielectric resonator oscillators (DROs). This resonator, which consists of a cooled sapphire ring or cylinder surrounded by a metallic shield, is capable of higher Q's than any other dielectric resonator, equaling those of quartz crystals at temperatures which can be reached by means of thermoelectric cooling [1]–[4]. At 10 to 20 K, it rivals the performance of superconducting resonators that require temperatures 10 times lower. This article reports on the results of tests on such a sapphire resonator at 9 to 10 GHz (X-band), which show Q's ranging from 3×10^8 at 77 K to 10^9 at 4.2 K.

The high Q's of these resonators depend not only on a reduction of losses internal to the sapphire but also on isolation of the resonant energy from losses in the surrounding metallic shield. With a dielectric constant (~10) only a fraction of that of other dielectric resonator materials, sapphire resonators are at a substantial disadvantage in this regard. This is overcome in the resonators of the present study through a process similar to the optical phenomenon of total internal reflection.

This article presents the results of both experiment and calculation, which show that effective isolation can be obtained in modes with 5 to 10 full waves around their perimeters. New computations for mode Q's and frequencies for high-mode numbers are presented on the basis of previously published solutions to the wave equation for an isolated isotropic dielectric sphere [8], the only finite geometry for which, to the author's knowledge, closed-form solutions have been devel-

oped. An approximate method is developed to allow calculations for right cylinders and for rings with rectangular cross sections. This method is based on solutions (also approximate) for a rectangular dielectric waveguide [9]. The ring is assumed to be just such a waveguide bent around on itself. Losses in the metallic shield are explicitly considered. The plethora of modes in the cylinder and sphere has led us to consider the ring resonator for further analysis and study. We find that an appropriate choice of ring dimensions can greatly increase the mode spacing without sacrificing the isolating properties of the mode.

Analyses of several different types of oscillator applications are presented. Possible applications include low noise microwave oscillators using only thermoelectric cooling and oscillators with both extremely low noise and high stability at temperatures of 77 K and below.

II. Background

Cryogenic sapphire resonators have been studied experimentally by Blair in Australia [1], [2] and by V. B. Braginsky and coworkers in the USSR [3], [4] with the aim of developing ultra-stable microwave oscillators and discriminators. Previous work has included measurements of mode frequencies and evanescent field decay lengths; measurement and calculation of the temperature and frequency dependence of the Q's; measurement of the fractional thermal coefficient of the resonant frequency; and development and study of stabilized oscillator performance.

In these experimental studies, sapphire losses are found to drop dramatically as the temperature is reduced below ambient, showing an approximately T^5 dependence for temperatures down to about 50 K, where a Q of approximately 10^8 is attained (for X-band). The loss mechanism responsible for this behavior has been identified by Gurevich [10] as phonon generation due to lattice anharmonicity. The T^5 dependence of the losses is predicted by this theory, as is a linear dependence on frequency. Both are borne out in experimental data, indicating that this source of loss is inherent in the sapphire and probably cannot be improved upon by better sample preparation. It seems appropriate, then, to use the currently observed high-temperature behavior as a basis with which to engineer filters and DROs.

The temperature dependence of the frequency of sapphire dielectric resonators has also been studied by both Blair and Braginsky *et al.* [1]–[4]. The fractional frequency variation with temperature $\partial F/\partial T/F$ is found to saturate at about 6×10^{-5}/K at high temperatures (>300 K), dropping as the coefficient of expansion "freezes out" at lower temperatures [1], [2]. It decreases to 3×10^{-6} at 77 K and falls as T^3 at lower temperatures to a value estimated to be 10^{-12}/K at a temperature of 1 K [3], [4]. The values found at 77 K and below could allow very impressive oscillator stability equivalent to that of quartz crystals at 40 K. At 10 K, the readily attainable temperature variation of 10 microdegrees would cause a fractional frequency variation of only 10^{-14}.

The stability demonstrated by oscillators using sapphire and sapphire-filled resonators shows the efficacy of this reduction in expansion coefficient. A frequency stability of 10^{-12} was demonstrated by the Russian group [3], [4] using a Gunn-excited oscillator, and stability better than 10^{-13} has been reported by the Australian group [1], [2] using a frequency-locked Gunn oscillator at room temperature. Using a sapphire resonator coated with superconducting lead, we have demonstrated stability better than 10^{-14} at 100 seconds. In the last case, the higher stability is not attributable to the superconducting coating but rather to the use of a ruby maser as the source of excitation [5]–[7].

While all of the oscillators just mentioned operate at temperatures below 2 K, the prospect of both high stability (due to the low expansion coefficient) and extremely low phase noise (due to high Q) in the temperature range from 10 K to 77 K is perhaps the most exciting aspect of their performance. Of great significance here are the relatively small and inexpensive cryocoolers available in this temperature range. In addition, comparison to conventional DROs and to cavity-stabilized microwave oscillators also indicates a dramatic reduction in phase noise using a sapphire resonator at approximately 170 K, a temperature achievable using thermoelectric cooling. Here the Q of 2×10^6 compares with values of 1 to 3×10^4 available from other microwave resonators, indicating a corresponding reduction in phase noise of 36 to 46 dB.

III. Isolated Modes in Dielectric Resonators

Isolated modes in dielectric resonators achieve weak coupling to the surrounding space not primarily by an impedance mismatch due to the large dielectric constant but rather by isolating properties of the mode itself. These modes can be understood from Fig. 1 as consisting of a wave trapped and slowed by a circular dielectric waveguide. The wave equation

$$k_r^2 + k_\theta^2 + k_z^2 = \epsilon(2\pi/\lambda)^2$$

allows a large value of k_θ inside the dielectric if the thickness and width of the ring are large enough to allow only small values of k_z and k_r, respectively. Outside the dielectric, however, the dielectric constant ϵ is 1, and this large value of k_θ, still required by the symmetry of the mode for some distance

outside the dielectric, requires an imaginary part in one of the other components (found in k_r) to satisfy this same wave equation. This region of evanescent, decaying fields forms a buffer between the waves in the dielectric and allows traveling waves farther out. These modes have been misnamed "whispering gallery" modes [3], [4] but are more properly seen as analogous to the phenomenon of total internal reflection in optics.

To the author's knowledge, solutions in closed form for the modes of cylindrically symmetric dielectric resonators are available only for the isotropic sphere and the infinite cylinder. Of these, the sphere, being a finite structure, is appropriate for consideration here. Following solutions published by Gastine *et al.* for the modes TE_{nmr} [8], we have calculated frequencies and Q's for $m = 1$, for $r = 1, 2$, and for n ranging up to relatively large values. These values are plotted in Fig. 2 and show an exponential increase in Q as n and the frequency are increased. A shortcoming in this calculation is the inability to account for the effect of a metallic shield, which is necessary to allow a reasonably small overall size as well as desirable to increase the radiation-limited Q's as shown in Fig. 2. It seems apparent that replacing the completely absorbing space surrounding the sphere by only slightly absorbing copper should improve the Q, but by how much? An upper limit would seem to be the product of the Q's; e.g., for $n = 7$ and $r = 1$, a radiation-limited Q of 3×10^4 (from Fig. 2) combined with a copper-can Q of 10^4 would indicate that Q's up to 3×10^8 might be possible, an attractive prospect. It also seems clear from Fig. 1 that the can must be in the evanescent region and that there would be some trade-off between isolation from can losses and overall size.

In order to test these ideas, we mounted an uncoated sapphire cylinder whose length and diameter were both approximately 5 cm inside a copper can large enough to provide a 1-cm gap at the outside and on the ends. At liquid-nitrogen temperature and below, we found modes with high Q ($Q > 10^7$) for frequencies above approximately 7.5 GHz. This frequency corresponds to $n = 8$ or 9 from Fig. 2 with a corresponding free-space radiation-limited Q of 10^5 to 10^6. Since the measured Q is higher than these values, some enhancement of the Q results from the low-loss properties of the shielding can.

However, the plethora of modes which we found gave us no hope of successfully identifying the modes on the basis of the spherical solutions. Furthermore, the prospect of oscillator design is daunting, given the existence of strongly coupled low-Q modes very near in frequency to weakly coupled high-Q modes.

A simple application of the wave equation to the geometry of Fig. 1, forcing a correspondence of k_r and k_z to half-wave solutions in the r and z directions, respectively, indicated that the number of modes might be greatly reduced without great penalty by a resonator with the geometry shown in Fig. 3. As a next step, and in order to obtain a more complete picture of the modes, we constructed a mode picture based on solutions for the modes of a rectangular dielectric waveguide derived by Marcatili [9], who identifies modes E^x_{pq} and E^y_{pq} with electric polarization in the x and y directions, respectively, and with p and q half waves in the x and y directions. Identifying the x, y, and z coordinates of these solutions with the r, z, and theta directions indicated in Fig. 3, identifying mode indices p and q with the mode multiplicity in the r and z directions, and introducing a mode number n corresponding to the number of full waves around the perimeter of the ring, we identify modes E^r_{npq} and E^z_{npq} for the ring.

Following Marcatili [9], we find E^y_{pq} modes for the rectangular dielectric waveguide as the solutions of $p^2X + q^2Y = 1$ where $X = (\pi/a)^2 \, (1 + 2A/\pi a)^{-2} \, (k_1^2 - k_\theta^2)^{-1}$ and $Y = (\pi/b)^2 \, (1 + 2A/\pi n^2 b)^{-2} \, (k_1^2 - k_\theta^2)^{-1}$ (where, in turn, $A = \lambda/2\sqrt{\epsilon - 1}$, $k_1 = 2\pi \, n/\lambda$, $n = \sqrt{(\epsilon)}$, $\lambda = c/f$, and a and b are the height and width of the ring cross section).

Explicitly accounting for the ring geometry by constraining the solution to exhibit n full waves around an effective ring perimeter r_{eff}, we define $k_\theta = 2\pi n/r_{\text{eff}}$, where r_{eff} is defined in terms of the inner and outer ring radii as $r^2_{\text{eff}} = (r_i^2 + r_o^2)/2$.

Solution of the wave equation outside the dielectric, matching the very large k_θ allowed inside, requires an imaginary part to at least one of the components of the wave vector k.

Decaying fields (imaginary components to the wave vector k) are required in the space just outside the dielectric by the wave equation as a result of the large value of k_θ allowed by the dielectric. A lower limit to the decay rate is obtained by identifying the decay length l_d as

$$l_d = 1/\pi \sqrt{(\pi/k_\theta)^2 - (2/\lambda)^2}$$

Assuming that the gap is much smaller than the radius, we identify the Q enhancement factor as the square of the field decay to the metallic wall a distance l_{gap} away:

$$\text{Q ratio} = \exp (2 \times l_{\text{gap}}/l_d)$$

We have calculated modes for a solid cylinder 5 cm in diameter and 5 cm long, identifying parameters $r_i = 0$, $r_o = 2.5$ cm, $r_{\text{gap}} = 1$ cm, $a = 2$ cm, and $b = r_o - r_i = 2.5$ cm, and for the ring in Fig. 3 with parameters given by $r_i = 1.5$ cm, $r_o = 2.5$ cm, $r_{\text{gap}} = 1$ cm, $a = 2$ cm, and $b = r_o - r_i = 1$ cm.

The results of these calculations are shown in Figs. 4 and 5. The predictions shown in Fig. 4 are in excellent qualitative agreement with the results of our measurements on the cylinder, confirming the validity of our approach. The efficacy of the ring geometry in reducing mode density is dramatically shown in a comparison of the two figures. The actual number of modes is larger than the number shown because modes with poor or no isolation are not shown. The calculations found 398 modes below 9 GHz for the cylinder and 60 for the ring. These modes are all doubly degenerate, a fact which was noted for many of them during the measurement process. Typical splitting of the modes was observed to be 10^{-5} to 10^{-6} fractional frequency deviation.

An inherent problem in the use of these resonators in active oscillators, and an important reason for choosing the ring for further study, is that the coupling of any mode to the external electronics will tend to scale in direct proportion to the coupling to the wall. This means that even though two modes may have very different Q's, if they are near to each other in frequency, mode selection may very well be a difficult problem. For example, if one of the modes is critically coupled to the active electronic elements, the other is likely to be nearly critically coupled as well.

IV. Q Measurements

Figure 6 shows the results of measurements of the Q of two of the modes of the 5 cm by 5 cm cylinder for temperatures below 77 K. Also plotted are higher temperature results reported by Braginsky *et al.* [3]. Good agreement is found with the results of these higher temperature data, confirming that these losses are inherent in the sapphire itself and are not due to impurities, surface treatment, etc. The leveling off of the loss reduction at about 10^{-8} is characteristic of the results reported by others and is probably due to impurities. The further Q improvement at the lowest temperatures is also typical, with the lowest point being marginally better than any others reported to date.

A consideration for resonator design is the requirement for surface finish and dimensional uniformity for the shaped dielectric cavity. Braginsky *et al.* [3] have used methods developed for optical fibers to estimate the losses caused by scattering from surface roughness. They find that for a resonator of centimeter dimensions, a roughness of 3-micrometer characteristic height will cause losses on the order of 10^{-10}. Although this value is smaller than any losses measured so far, the resonator used in the measurements reported here was fabricated with an optical quality polish on all surfaces to assure no loss contributions from this source. Precautions such as acid etch and purified alcohol rinse were taken to assure that no foreign material adhered to the surface.

V. Predictions of Oscillator Noise Performance

The reduction in phase noise over that in conventional DRO and cavity oscillators which would result from the high Q of a cooled sapphire resonator is shown in Fig. 7. Q's of 10,000 and 30,000 are assumed for the conventional oscillators, respectively, along with values from Fig. 6 for the sapphire DROs. Also shown is a further reduction which would result from the application of ruby maser technology to such oscillators.

Multiplicative $1/f$ noise $S_\phi(f)$ in the active device is assumed to be −100 dBc/f (/Hz) [11], [12] and −130 dBc/f (/Hz) for the curves indicating maser excitation.[1] The latter value corresponds to an upper limit obtained in tests of a low-Q S-band ruby maser oscillator [5], a value substantially quieter than that reported for any other active microwave device. It has been well documented that multiplicative $1/f$ noise in semiconducting devices can be reduced by operating devices in parallel or, similarly, by large gate dimensions. Thus it seems likely that the low $1/f$ noise in the ruby is due to its very large volume (~1 cm^3). Ruby masers have been operated at temperatures as high as 90 K and at frequencies up to 42 GHz [11], [12]. Their application to low noise oscillators could open a new window in low noise oscillator capability.

[1]This compares to the best X-band GaAs FET multiplicative noise (−109 dBc/f [Hz]) thus far discovered by the author.

References

[1] D. G. Blair and S. K. Jones, "High-Q Sapphire-Loaded Superconducting Cavities and Application to Ultrastable Clocks," *IEEE Trans. Magnetics*, vol. MAG-21, p. 142, 1985.

[2] D. G. Blair and I. N. Evans, "High-Q Microwave Properties of a Sapphire Ring Resonator," *J. Phys. D: Appl. Phys.*, vol. 15, pp. 1651–1656, 1982.

[3] V. B. Braginsky, V. P. Mitrofanov, and V. I. Panov, *Systems with Small Dissipation*, Chicago: University of Chicago Press, pp. 85–89, 1985.

[4] V. I. Panov and P. R. Stankov, "Stabilization of Oscillators with High-Q Leucosapphire Dielectric Resonators," *Radiotekhnika i Electronika*, vol. 31, p. 213, 1986.

[5] D. M. Strayer, G. J. Dick, and J. E. Mercereau, "Performance of a Superconducting Cavity of Superior Quality," *IEEE Trans. Magnetics*, vol. MAG-22, 1986.

[6] S. Thakoor, D. M. Strayer, G. J. Dick, and J. E. Mercereau, "A Lead-on-Sapphire Superconducting Cavity of Superior Quality," *J. Appl. Phys.*, vol. 59, p. 854, 1986.

[7] G. J. Dick and D. M. Strayer, "Development of the Superconducting Cavity Maser as a Stable Frequency Source," in *Proceedings of the 38th Annual Frequency Control Symposium*, p. 414, 1984.

[8] M. Gastine, L. Courtois, and J. L. Dormann, "Electromagnetic Resonances of Free Dielectric Spheres," *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-15, p. 694, 1967.

[9] E. A. J. Marcatili, "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," *Bell Systems Technical Journal*, vol. 48, p. 2071, 1969.

[10] V. L. Gurevich, *Kinetics of Phonon Systems*, Moscow, 1980.

[11] H. Reitbock and A. Redhardt, "A Molecular Amplifier for a Working Temperature of 90 K" (in German), *Naturforsch.*, vol. 17A, p. 187, 1962.

[12] C. Moore and D. Neff, "Experimental Evaluation of a Ruby Maser at 43 GHz," *IEEE Trans.*, vol. MTT-30, p. 2013, 1982.

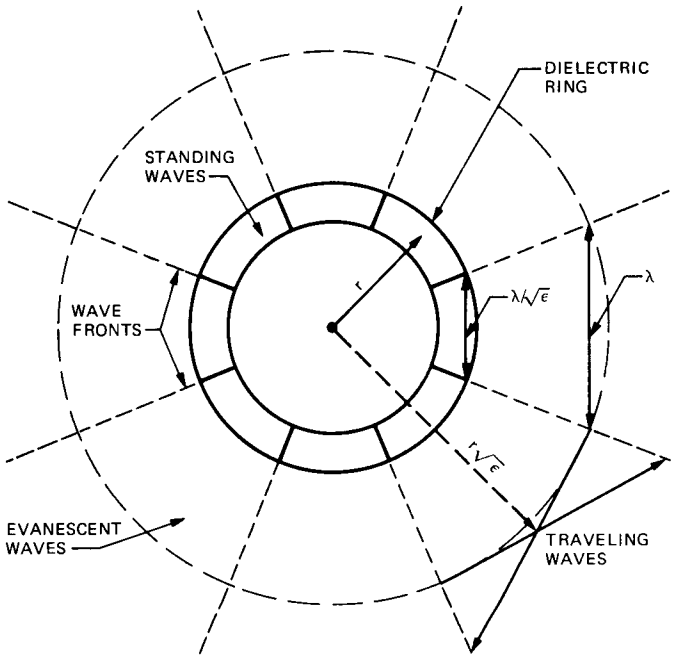


Fig. 1. Diagram showing the character of the electromagnetic field in the vicinity of a dielectric ring for an eight-fold cylindrically symmetric mode

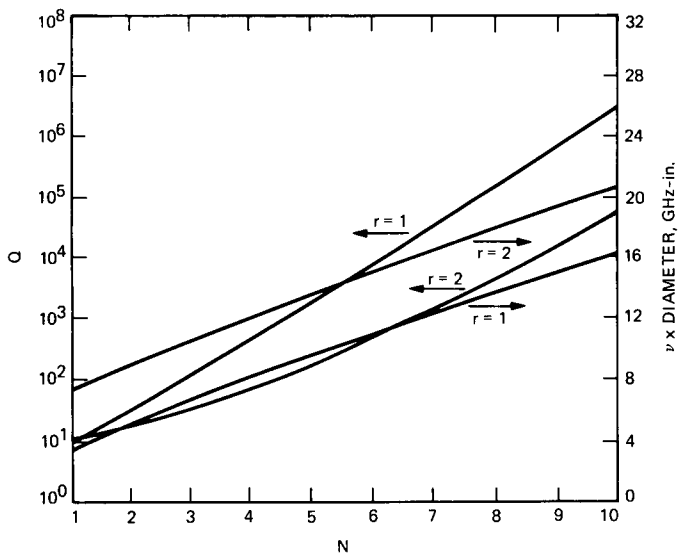


Fig. 2. Radiation-limited Q and frequency for TE_{nmr} modes of an isolated sphere with epsilon = 10, m = 1, and r = 1, 2

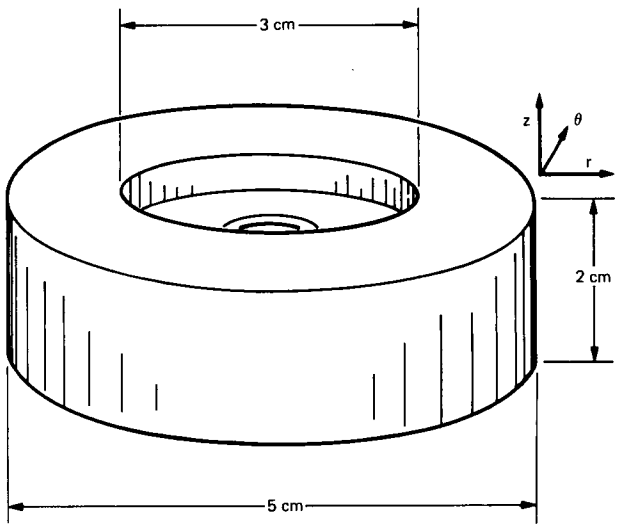


Fig. 3. Sapphire ring construction showing directional axis identification at ring perimeter

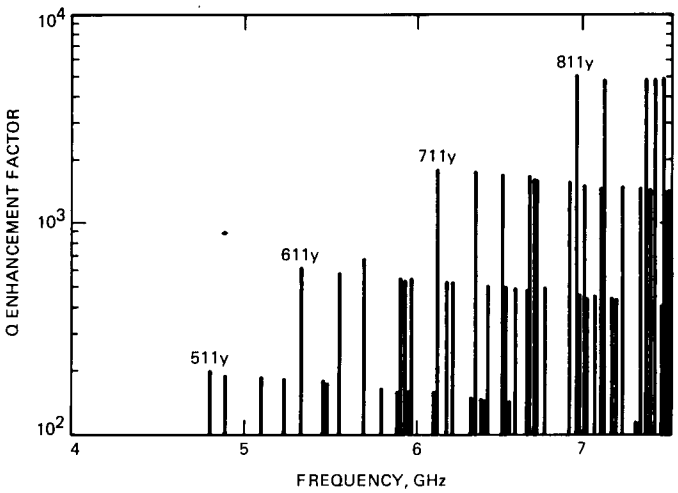


Fig. 4. Calculated mode frequencies and Q enhancement factors for a dielectric sapphire cylinder 5 cm in diameter and 5 cm high surrounded by a lossy shield 1 cm from the surface

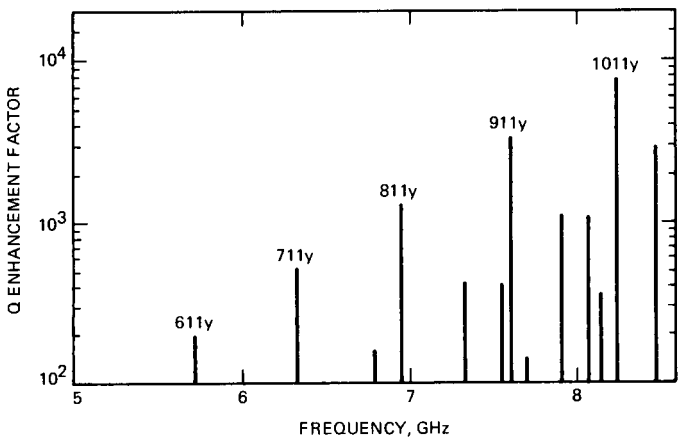


Fig. 5. Calculated mode frequencies and Q enhancement factors for the ring shown in Fig. 3 surrounded by a lossy shield 1 cm from the surface

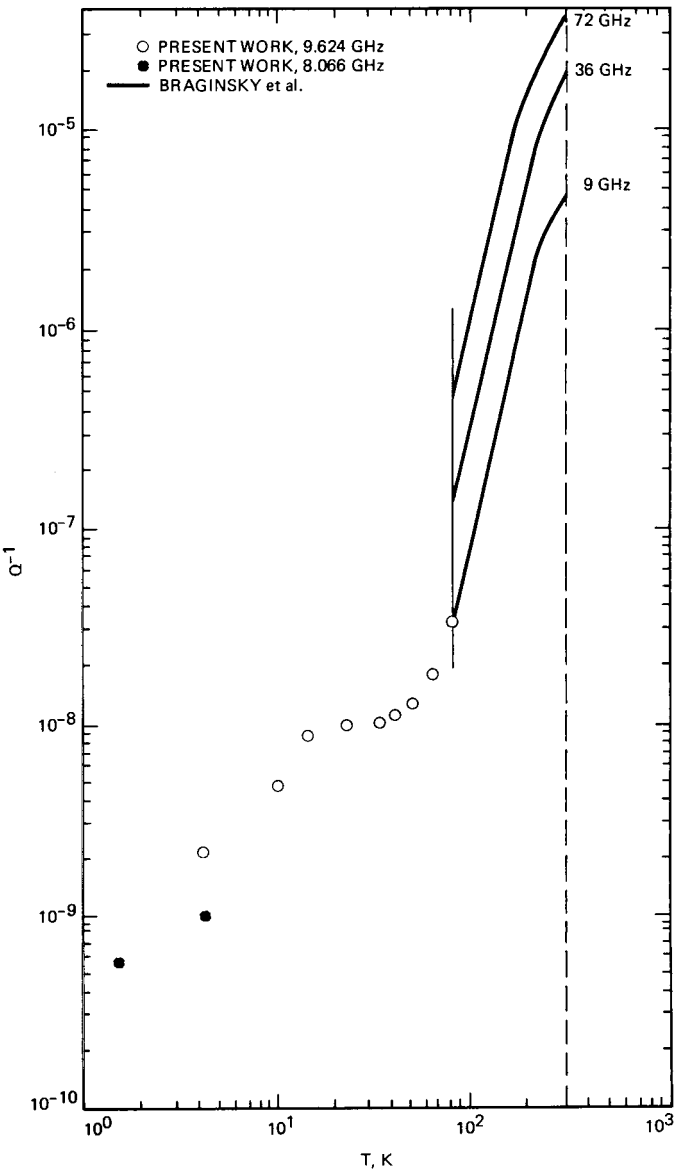


Fig. 6. Q measurements for a sapphire cylinder 5 cm in diameter and 5 cm high contained in a lead-plated shielding can approximately 1 cm away (also shown are higher-temperature data by Braginsky *et al.* [2]

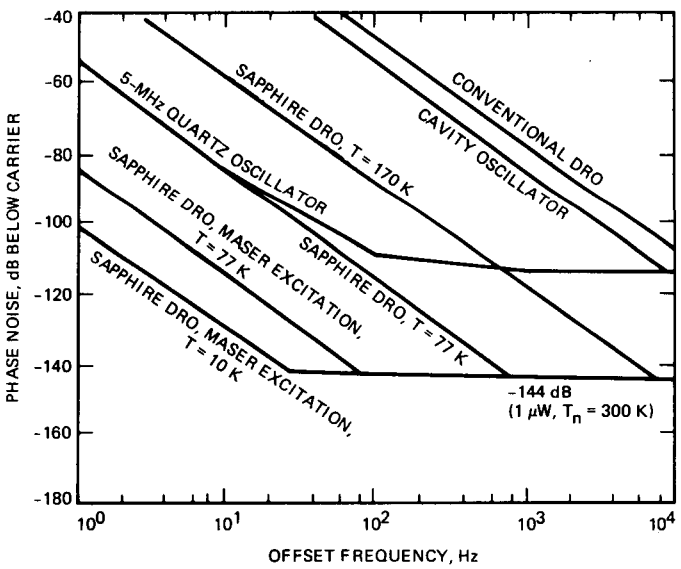


Fig. 7. Phase noise for various X-band sources, including conventional DRO and cavity oscillators, a state-of-the-art quartz crystal oscillator referenced to 10 GHz, and predictions for several sapphire DROs

Inductance Effects in the High-Power Transmitter Crowbar System

J. Daeges and A. Bhanji
Radio Frequency and Microwave Subsystems Section

The effective protection of a klystron in a high-power transmitter requires the diversion of all stored energy in the protected circuit through an alternate low-impedance path, the crowbar, such that less than 1 joule of energy is "dumped" into the klystron during an internal arc. A scheme of adding a bypass inductor in the crowbar-protected circuit of the high-power transmitter was tested using computer simulations and actual measurements under a test load. Although this scheme has several benefits, including less power dissipation in the resistor, the tests show that the presence of inductance in the portion of the circuit to be protected severely hampers effective crowbar operation.

I. Introduction

Modern-day high-power transmitters use large and expensive components. For example, a klystron may cost as much as $250,000. The protection of these high-cost items from destructive arcs and overloads is therefore of paramount importance in terms of economy, component life, and reliable operation of the transmitter.

One of the special protection devices in a high-power transmitter is the crowbar unit. When an internal arc in the klystron is sensed, the crowbar is fired within 5 to 10 microseconds, thereby diverting the large stored energy in the power supply system and preventing the disaster that would result if more than 1 joule of the energy were "dumped" into the arc.

In the present DSN transmitter crowbar design (Case 1), a series isolation resistor in the protected portion of the crowbar circuit dissipates power in normal steady-state operation. The disadvantages of this system are that the dissipated power is lost, reducing power to the klystron, and that the series resistor design requires a large amount of physical space. A new scheme was developed (Case 2) that would add a large bypass inductor in parallel with the resistor. DC steady-state current would then flow through the inductor, preventing power dissipation in the resistor. The resistor value was subsequently increased to match the characteristic impedance of the cable.

This article presents data taken from computer simulation and physical measurements under a test load of both the Case 1 and Case 2 schemes (see Fig. 1). The data shows that the presence of inductance in the portion of the system to be protected (klystron/crowbar discharge loop) severely limits the effectiveness of the crowbar.

II. Crowbar Operation and Simulation

A. Crowbar Operation

Observations and simulations were conducted on the DSS-13 S-band klystron high-power transmitter crowbar system to determine the effectiveness of the crowbar when inductance is present in the protected portion of the circuit.

Figure 1 is a simplified schematic of the system that is useful for analyzing various transient loop currents and voltages. The symbols used are defined as follows:

i_f	Fault current flowing through the "1-joule crowbar test wire" when it is shorted to simulate an arc fault
i_c	Crowbar current
E	Power supply voltage
R_F	Discharge resistor
R_1	Crowbar resistor
R_2	Crowbar isolation resistor
L_2 and L_3	Lumped and stray inductance in the crowbar and protected circuit
R_{2M}	Matching impedance to the HV cable
L_{2M}	Steady-state bypass inductance
L_1	Filter inductor
C_1	Filter capacitor
C_2	Energy storage capacitor
C_s	Stray capacitance in protected circuit
r_a	Resistance of the "1-joule crowbar test wire" (this wire is 2.237-inch-long #36 AWG soft copper, which will fuse when 1 joule of energy is passed through it)
t_0	Time at initiation of fault (switch S_a is closed)
t_1	Time when crowbar fires (switch S_a is closed)
t_2	Time when power supply circuit breaker opens (switch S_1 is opened)
Z_o	Characteristic impedance of HV cable
τ	One transit time delay of the cable
R_L	Klystron load equivalent impedance

To make a meaningful analysis of the crowbar system, actual values and, in some cases, estimated values were assigned to the circuit of Fig. 1 as follows:

$E = 40$ kV

$R_1 = 1$ ohm

$R_F = 39$ ohms

$C_1 = 0.2\ \mu$F

$L_1 = 1$ H

$C_2 = 1\ \mu$F

$L_2 = 4\ \mu$H (estimated stray inductance of the wire)

$L_3 = 6\ \mu$H (estimated stray inductance of the wire)

$Z_o = 46$ ohms

$\tau = 1\ \mu$s

$R_L = 4$ kohms (estimated klystron normal operating impedance)

$C_s = 1000$ pF estimated stray capacitance

$r_a = 0.077$ ohm

B. Test Cases

Measurements were taken for the following two operational scenarios:

Case 1: Minimal stray inductance in the protected circuit with $R_2 = 10$ ohms.

Case 2: Large bypass inductor, $L_{2M} = 4$ mH inductance; $R_{2M} = 40$ ohms.

C. Crowbar Measurements

The crowbar and the arc currents were measured using Pearson Model 101 current probes. The currents were recorded using an HP 1631AD logic analyzer. To simulate an arc, the "1-joule test wire" was shorted to ground, paralleling the 4-kohm equivalent klystron resistive load. The crowbar ignitron switch was fired 10 microseconds later.

D. Computer Simulation

SPICE,[1] a general purpose circuit simulation program for nonlinear dc, nonlinear transient, and linear ac analysis, was used to analyze the crowbar model.

For the purpose of simulating a fault condition (arc), S_a is closed at some time t_0, paralleling the klystron load with a low resistance path ("1-joule test wire"). At time t_1, 10 microseconds later, switch S_c is closed to simulate the crowbar action. Although the trigger pulse is 150 microseconds, the switch S_c can cease conduction early if the voltage reverses across it during the crowbar discharge (underdamped crowbar discharge loop) or if current through it falls below the value necessary for continued conduction (normally less than 10 amps).

III. Results

SPICE-simulated waveforms of arc current through "1-joule test wire" and crowbar current are shown in Fig. 2

[1]SPICE Version 2.0, Intusoft, P.O. Box 6607, San Pedro, CA 90734.

for Case 1. Figure 2a shows a time scale of 50 microseconds, and Fig. 2b shows a time scale of 1 millisecond.

Physically measured results for Case 1 are shown in Fig. 3. The results are plotted using the same time scales as the SPICE-simulated waveforms in Fig. 2.

SPICE-simulated waveforms for Case 2 are shown in Fig. 4 with time scales of 50 microseconds (Fig. 4a) and 1 millisecond (Fig. 4b).

Physically measured results for Case 2 are shown in Fig. 5. The results are plotted using the same time scales as the SPICE-simulated waveforms in Fig. 4.

From the figures it can be noted that:

(1) Experimental results closely match the computed results.

(2) For Case 1, the arc current decays to zero within 5 microseconds after the crowbar fires (Figs. 2a and 3a), and the crowbar current initially builds up as the ignitron (switch S_c) starts conducting and diverting the energy in the high-voltage cable and in C_2. As C_2 discharges, the crowbar current starts decaying. About 300 microseconds later, the stored energy in L_1 and the follow-on current from the power supply kick in, and the crowbar current starts rising again (Figs. 2b and 3b) until switch S_1 disconnects the power supply.

(3) For Case 2, the arc current drops initially as soon as the crowbar is fired. However, the stored energy in L_{2M} sustains the arc current at a much lower value (Figs. 4a and 5a). Unfortunately, the presence of this large L_{2M} in the protected circuit produces an underdamped voltage condition in the ignitron (Fig. 6) and turns it off. Rather than flowing through the crowbar, which is now turned off, the undissipated energy in C_2 and the inductive energy in L_1 continue to flow through the arc, building the current (Figs. 4b and 5b) until the power supply is turned off.

IV. Discussion

In most applications, the actual energy that a klystron can safely dissipate during an internal arc is an unknown quantity. However, the consensus in the industry is that this dissipated energy should be no more than 1 joule.

The crowbar circuit is tested by substituting a "1-joule test wire" for the klystron. The test wire is purposely shorted through switch S_a to simulate an arc. The energy dumped in this wire is then the resistance of this wire times the square of the arc current times the length of time the arc current continues to flow.

When the crowbar was tested for both Case 1 and Case 2, the 1-joule wire survived in Case 1, indicating less than 1 joule of energy in the arc. The wire was completely fused (evaporated) in Case 2, indicating more than 1 joule in the arc.

In Case 2, as discussed above, the energy stored in L_{2M} and its actual inductance can be considered the driving force that opposes the quenching of the arc. At the same time, the reverse voltage turns off the crowbar, thus adding the rest of the energy from the power supply system into the arc.

V. Conclusions

The information obtained from the tests and computer simulations described above gives conclusive proof that inductance in the protected portion of the circuit severely limits the effectiveness of the crowbar. The inductance increases the total charge that flows through the arc (fault) and impedes the quenching of the arc. Therefore, inductance in the protected portion of the circuit should be kept to the minimum dictated by the physical layout.

Also, the total energy that flows into the fault will be reduced if the delay time to fire the crowbar is kept to a minimum, preferably less than 1 microsecond. Increasing the R_2 (isolation) series resistance would help reduce the energy dissipated in the arc. It would also keep the discharge loop either critically damped or overdamped, thus preventing the ignitron from shutting off. However, R_2 cannot be increased beyond 5 to 10 ohms, since these elements dissipate power during normal klystron operation, resulting in reduced efficiency of the system.

Finally, a note on computer simulation. A crowbar system is a large and expensive piece of equipment. It is not easily modeled on the laboratory bench with smaller components, and yet it must be modeled and more fully understood to ensure reliable operation. Since these simulations and experimental measurements agree quite well, we now have a tool that allows us to try out new concepts rapidly by giving us considerable insight into how the circuit will behave before expensive components are built.

C-2

Acknowledgment

The authors wish to express their appreciation to Keith Gwen, who performed various modifications to the crowbar system and assisted in taking the measurements.

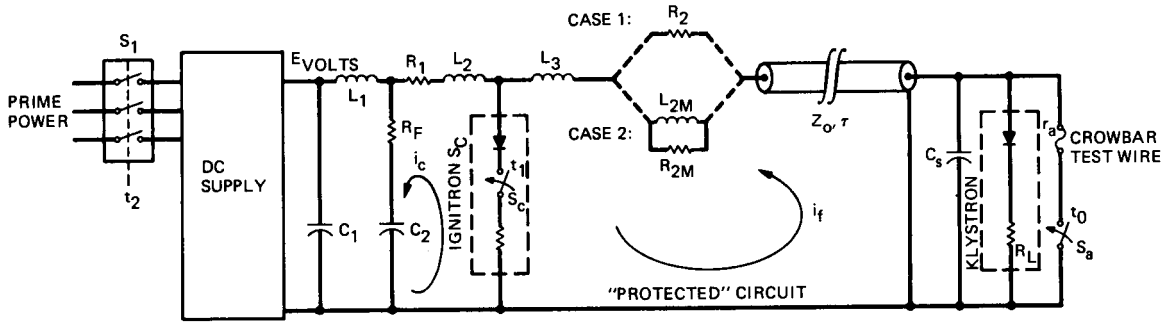


Fig. 1. Crowbar system, simplified schematic diagram

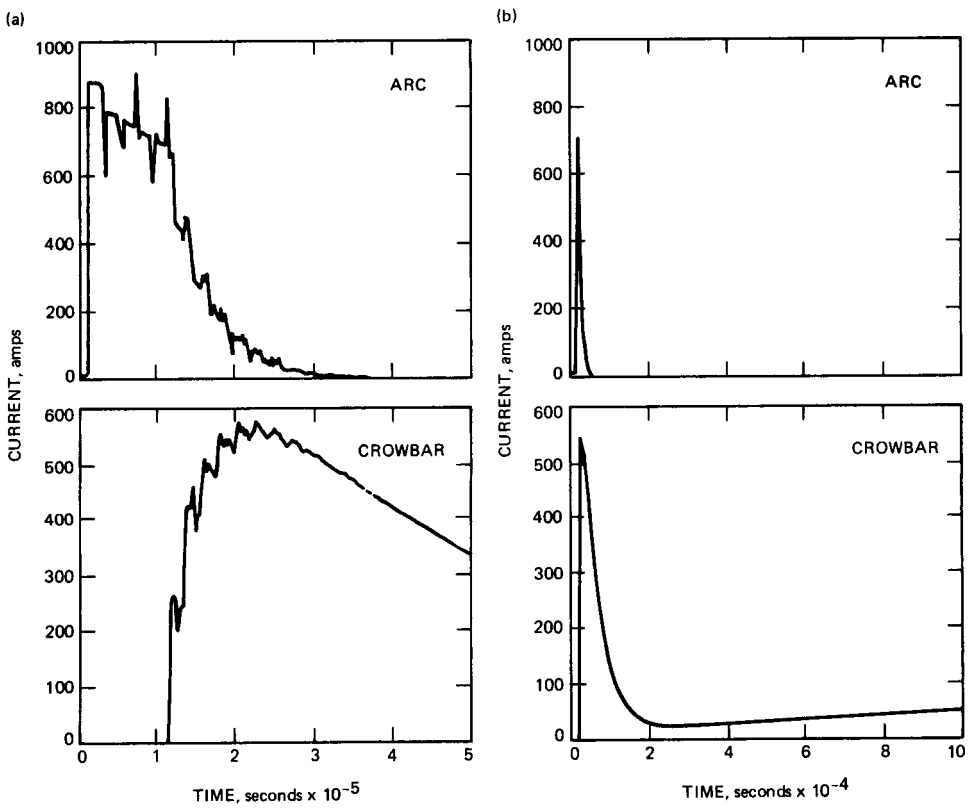


Fig. 2. SPICE-simulated waveforms for Case 1: (a) 50-microsecond time scale; (b) 1-millisecond time scale

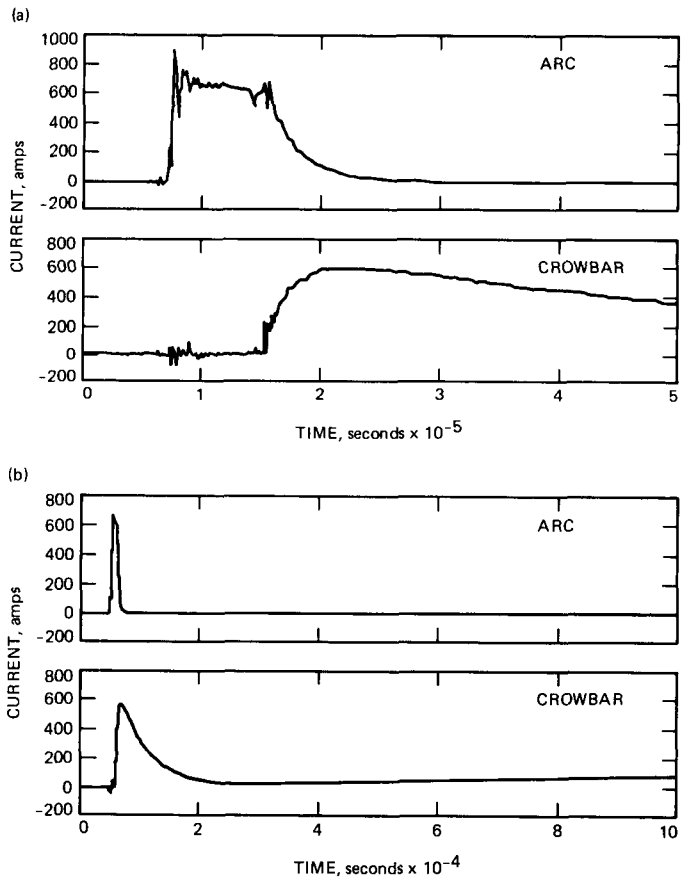


Fig. 3. Physically measured results for Case 1: (a) 50-microsecond time scale; (b) 1-millisecond time scale

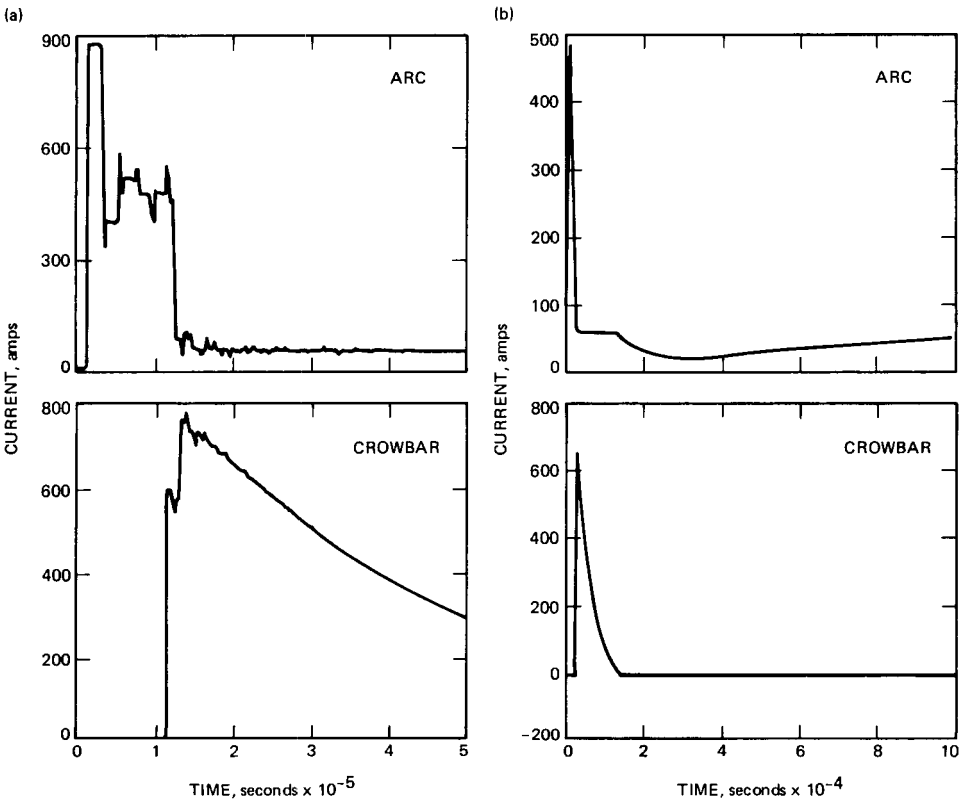


Fig. 4. SPICE-simulated waveforms for Case 2: (a) 50-microsecond time scale; (b) 1-millisecond time scale

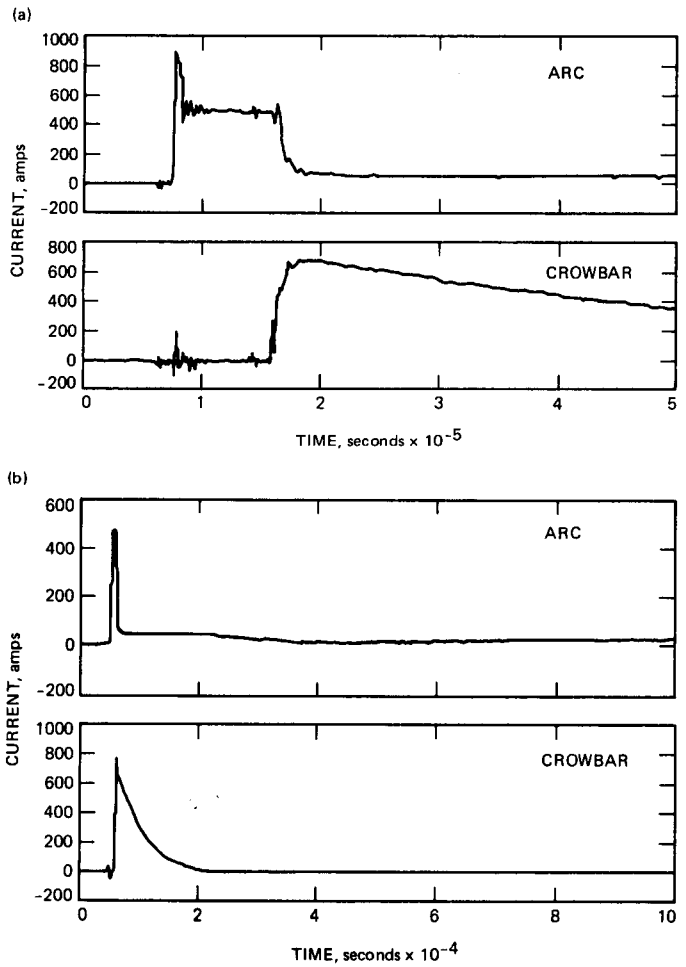


Fig. 5. Physically measured results for Case 2: (a) 50-microsecond time scale; (b) 1-millisecond time scale

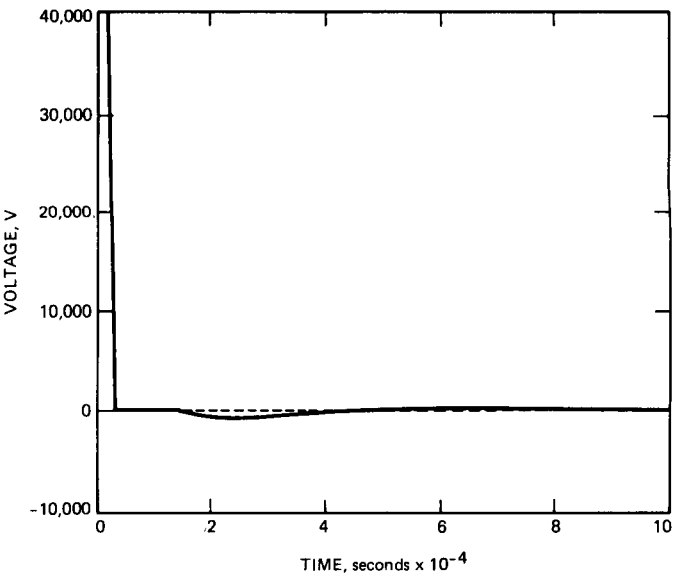


Fig. 6. SPICE-simulated waveform of the crowbar ignitron voltage

Low-Noise Cryogenic Transmission Line

D. Norris
Radio Frequency and Microwave Subsystems Section

New low-noise cryogenic input transmission lines have been developed for the DSN at 1.668 GHz and 2.25 GHz for cryogenically cooled Field Effect Transistor (FET) and High Electron Mobility Transistor (HEMT) amplifiers. These amplifiers exhibit very low noise temperatures of 5 K to 15 K, making the requirements for a low-noise input transmission line critical. Noise contribution to the total amplifier system from the low-noise line is less than 0.5 K for both the 1.668-GHz and 2.25-GHz FET systems. The 1.668-GHz input line was installed in six FET systems which were implemented in the DSN for the Venus Balloon Experiment. The 2.25-GHz input line has been implemented in three FET systems for the DSN 34-m HEF antennas, and the design is currently being considered for use at higher frequencies.

I. Introduction

The purpose of the cryogenic input transmission line is to direct a desired RF signal from the room-temperature environment to the input of a cryogenically cooled amplifier while adding as little noise as possible. This article describes the design and performance of a cryogenically cooled coaxial probe transmission line that adds less than 0.1 K to the noise temperature of a maser amplifier and less than 0.5 K to the noise temperature of a cryogenic transistor amplifier. This input line was first designed and implemented in 1973 [1] for 2.3-GHz traveling-wave maser (TWM)/closed-cycle refrigerator (CCR) systems, which operate at a physical temperature of 4.5 K. More recently, the line was redesigned for 1.668-GHz and 2.25-GHz FET/CCR and HEMT/CCR systems, which operate at a 12-K nominal physical temperature. The FET amplifier, as well as the similar high electron mobility transistor (HEMT) amplifier, has a much wider bandwidth than masers, and the coaxial probe unit was successfully broadbanded for 1.668- and 2.25-GHz use as a result of this design effort.

II. Design

Figure 1 shows a cutaway view of the low-noise cryogenic transmission line assembly. Figure 2 is a photograph of the input transmission line installed on a TWM/CCR assembly. Figure 3 is a photograph of the components. The RF input signal comes into the WR 430 waveguide flange and is coupled to a coaxial probe. A quartz dome window in the waveguide establishes a vacuum environment for the entire probe/center conductor, which is cantilever-supported from the cold (4.5-K or 12-K) end. The entire length of the probe/coaxial center conductor can be cooled to the CCR first-stage temperature. Since a major portion of the insertion loss and noise of a

coaxial line is contributed by the center conductor's resistivity and physical temperature, this approach provides very low noise contribution.

A waveguide shorting plate is placed at one end of the waveguide approximately 1/4 wavelength behind the probe. The outer coaxial conductor is thin-wall SS tubing plated in the inside surface with a few skin depths of copper (120 microinches nominal). This copper thickness is adequate for RF surface conductivity but does not significantly degrade the thermal insulation provided by the SS tubing, which is tied to room temperature at one end, to the 70-K heat station, and to the 4.5-K (or 12-K in the case of HEMT coolers) final heat station in the CCR.

The impedance of the coaxial line was chosen to be 77 ohms for minimum loss (as described in [2]) and is transformed to 50-ohm output impedance at the SMA connector with a step transformer.

Figure 4 shows a Smith chart with sample data of the tuning process (as described in [3]) that was followed to optimize both the distance from the probe to the waveguide shorting plate and the length of the probe within the WR 430 waveguide cavity. This was done at a single frequency representing the desired band center. The four curves were obtained for the four labeled probe-to-shorting plate dimensions, and each data point on each curve represents a different probe length in 0.51-cm increments. In this manner, the optimum dimensions can be obtained by interpolation of the data for point A. In the case of the 2.25-GHz assembly, the optimum dimensions were found to be 3.048 cm for the probe length and 3.302 cm for the probe-to-back short. Increased bandwidth was obtained by the addition of shunt capacitance near the SMA connector.

III. Performance

The measured return loss of the assembled transmission line (shown in Fig. 5) is greater than 20 dB between 2.00 GHz and 2.60 GHz. The heat load on a three-stage, 4.5-K CCR is less than 100 milliwatts. The contribution to the input noise temperature of a 2.3-GHz maser amplifier (in a 4.5-K CCR) is estimated to be less than 0.1 K. The contribution to the noise temperature of the 2.3-GHz HEMT and FET amplifiers (in a 12-K CCR) is less than 0.5 K.

IV. Conclusions

The 1.668-GHz low-noise cryogenic transmission line input assembly was implemented for the first time in six FET/CCR systems in support of the Venus Balloon Experiment. The 2.25-GHz input is also being used in the Deep Space Network in a HEMT/CCR at DSS-13 and in three FET/CCRs at DSS-15, DSS-45, and DSS-65. These transmission lines provide sufficient bandwidth to be useful with the bandwidths achieved by state-of-the-art cryogenic FET and HEMT amplifiers. This design is now being considered for use at higher frequencies as a more compact alternative to waveguide input transmission lines.

References

[1] R. Clauss and E. Wiebe, *JPL Technical Report 32-1526*, vol. XIX, February 15, 1974.

[2] S. F. Adam, *Microwave Theory and Applications*, New York: Prentice-Hall, pp. 37–41, 1969.

[3] G. L. Ragan, "Microwave Transmission Circuits," MIT Radiation Laboratory Series, New York: McGraw-Hill, pp. 318–322, 1948.

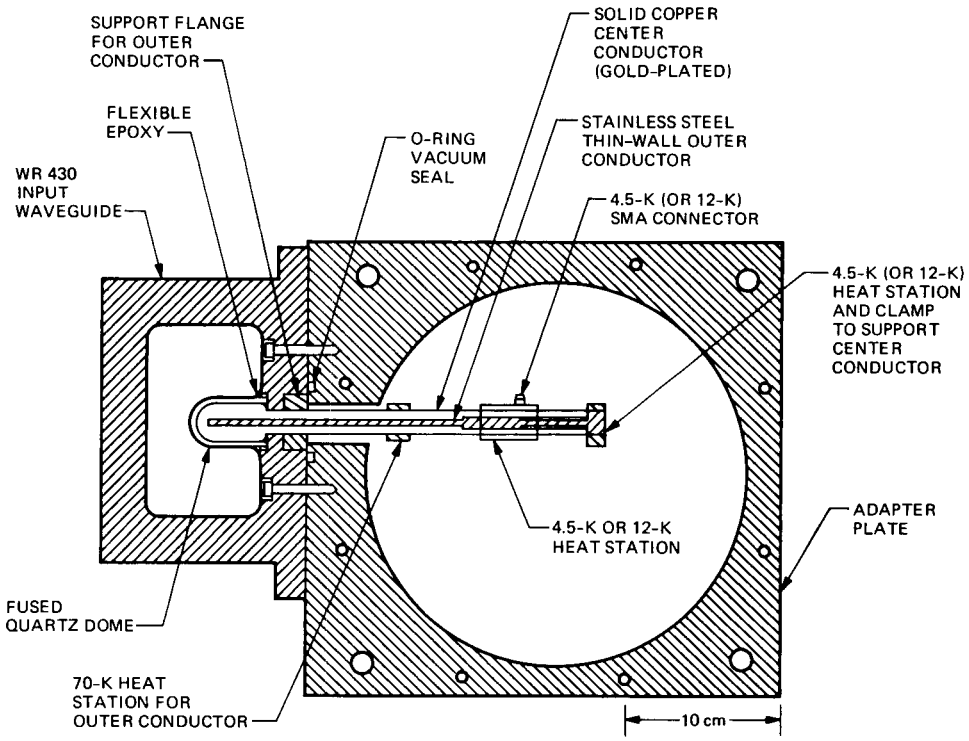


Fig. 1. Cutaway view of cryogenic input transmission line assembly maser configuration (FET and HEMT applications are similar)

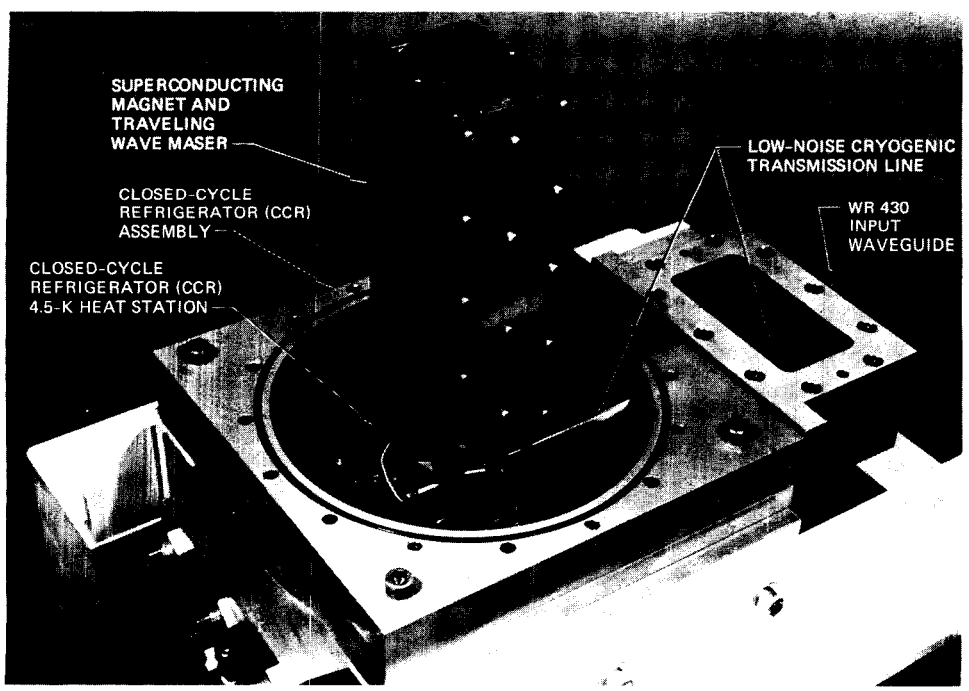


Fig. 2. A 2.3-GHz traveling-wave maser (vacuum housing and radiation shields are partially removed)

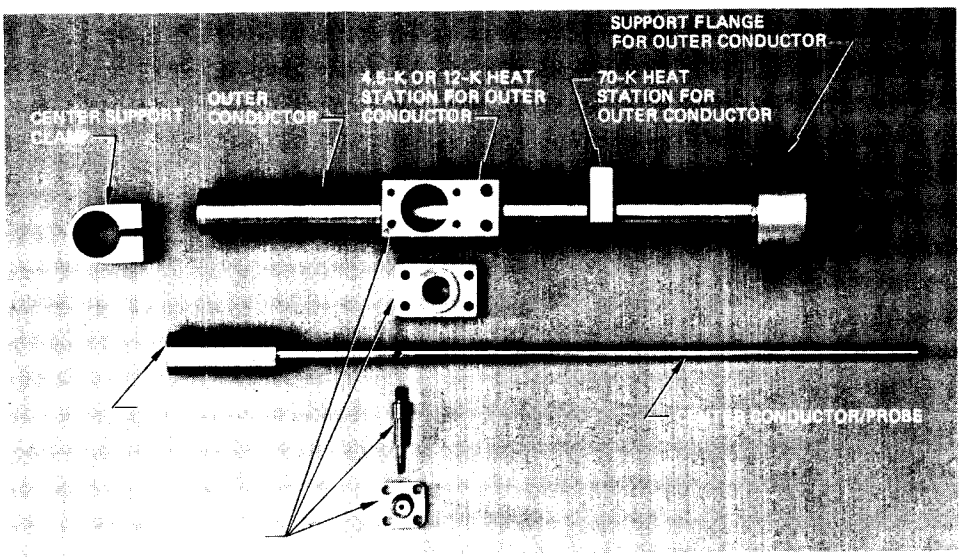


Fig. 3. Components of cryogenic input transmission line

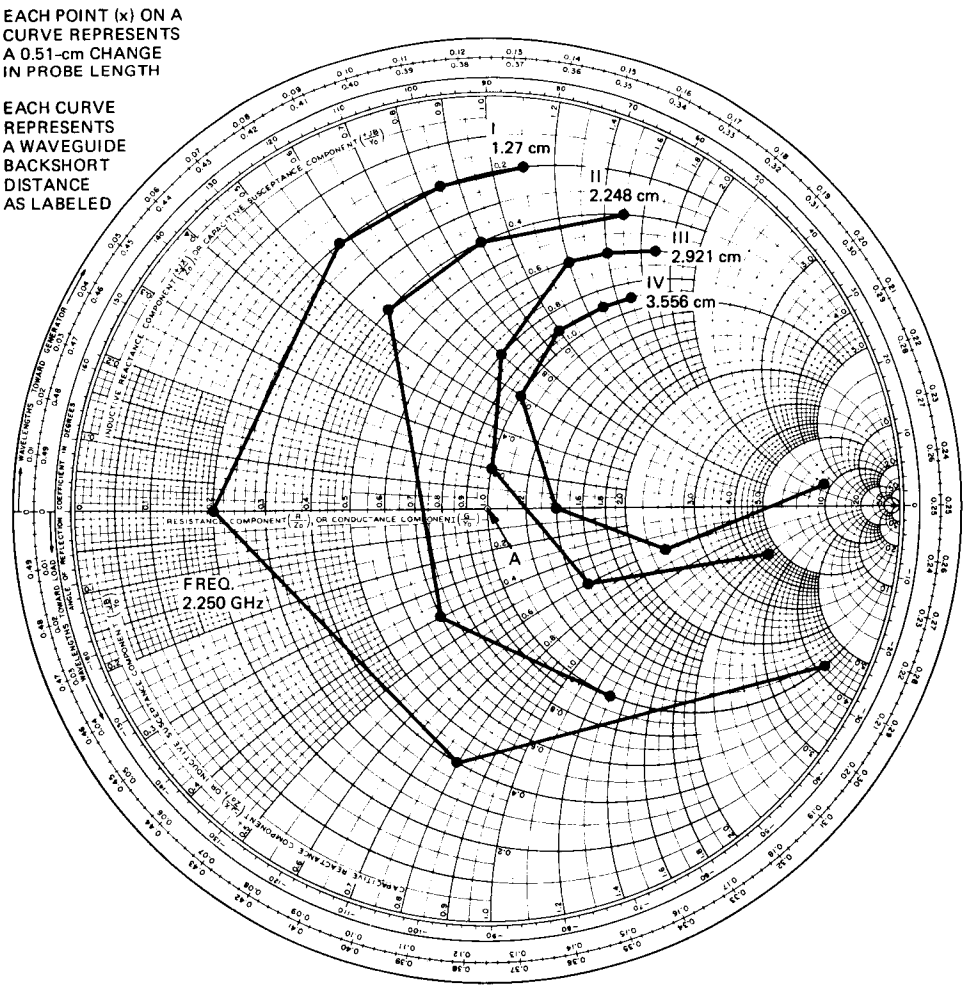


Fig. 4. Smith-chart plot of 2.25-GHz transmission line assembly S_{11}

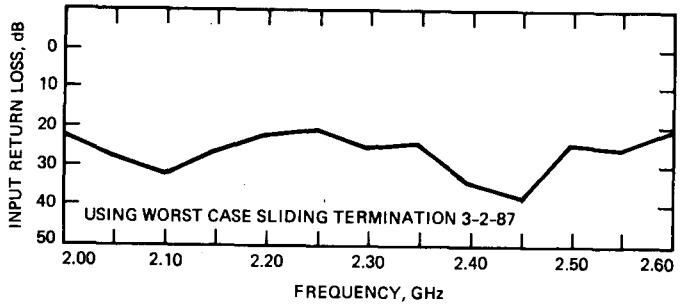


Fig. 5. Return loss versus frequency at 2.25 GHz

A 2.3-GHz Cryogenically Cooled HEMT Amplifier for DSS 13

L. Tanida
Radio Frequency and Microwave Subsystems Section

A prototype 2.3-GHz (S-band) high electron mobility transistor (HEMT) amplifier/closed-cycle refrigerator (CCR) system was installed in the DSS-13 feedcone in August 1986, replacing the 2.3-GHz maser. The amplifier is cryogenically cooled to a physical temperature of 12 K and provides 31.5 K antenna system noise temperature and 29 dB of gain. The HEMT device used in the amplifier is a prototype developed for JPL under an R&D contract with the General Electric Company.

I. Introduction

Masers are the lowest noise amplifiers known. They have been used in the Deep Space Network for over 25 years for spacecraft downlink communication and navigation. Their 4.5-K cooling requirements and their complexity make them costly to implement and to maintain. Cryogenically cooled high electron mobility transistor (HEMT) amplifiers [1] are approaching the noise performance of the maser in the 1- to 4-GHz region at one-fifth the cost. This report describes the design, assembly, and performance of the first HEMT/CCR system designed for the DSN. It operates over an instantaneous bandwidth of 2220 to 2330 MHz (limited by an input bandpass filter) with a nominal input noise temperature of 8.7 K at the room temperature waveguide input flange (Table 1). Antenna system noise temperature at 2295 MHz was 31.5 K in comparison to the maser system noise temperature of 28.7 K (Table 2). This higher system noise temperature is offset by significant cost savings, greater reliability, and a wider bandwidth capability useful for VLBI experiments.

II. General Description

Figure 1 shows a block diagram of the DSS-13 2.3-GHz microwave front end, including the 2.3-GHz HEMT/CCR system. Figure 2 shows a photograph of the HEMT/CCR package with vacuum housing and radiation shield removed to show the components mounted on the 12-K refrigerator station. The 12-K closed-cycle refrigerator contains a cryogenically cooled input transmission line, a cryogenic bandpass filter, a cryogenic isolator, and a three-stage amplifier which contains a HEMT in the first stage and GaAs FET devices in the second and third stages.

III. Detailed Description

A. Cryogenic Input Transmission Line

The function of the input transmission line is to transmit the input RF signal from the room temperature WR 430 waveguide flange to the cryogenic low-noise amplifier while adding the least noise possible, adding an acceptable amount of heat load to the CCR, and remaining within the size constraints of the refrigerator. The input signal is converted from waveguide to a coaxial line with a copper center conductor that is cooled along its entire length to a temperature of 12 K and with an outer conductor (0.5-inch OD, 0.010-inch-wall stainless steel) that varies in temperature from 300 K (room temperature) to 12 K along its length. A quartz dome vacuum window is epoxied into the WR 430 waveguide/coaxial transition so that

the entire coaxial center conductor (including probe section in the WR 430 waveguide) and outer conductor are in a thermally insulating vacuum environment. This assembly has been used previously in Block IV and Block V traveling-wave masers and is described in [2]. The noise contribution of this assembly is estimated to be less than 0.5 K.

B. Cryogenic Bandpass Filter

HEMT and FET amplifiers are inherently very broadband devices. The amplifier in this system has over 10-dB gain from 1.0 to 3.4 GHz. Therefore, it was deemed necessary to include a bandpass filter to protect the system from out-of-band radio frequency interference (RFI) and from transmitter leakage when operating in a diplexed mode on the antenna. The filter is an evanescent mode, six-section unit supplied by K&L Microwave. This particular design was selected among other designs for the following reasons: (1) its small size permitted convenient mounting on the 12-K stage of the refrigerator; (2) the frequency shift from room temperature to 12 K is minimal (7.5 MHz); and (3) attenuation at the transmitter frequency (2120 MHz) is -55 dB. The filter is made of aluminum and is silver-plated to give the lowest microwave surface resistivity at 12 K. The noise contribution of this assembly is calculated to be approximately 0.9 K.

C. Cryogenic Isolator

A cryogenically cooled isolator is included at the input of the HEMT amplifier to provide good input match over the entire bandwidth. The cryogenic isolator, supplied by Passive Microwave Technologies, has a return loss of greater than 15 dB at 2200 to 2330 MHz and an insertion loss of 0.3 dB. The estimated noise contribution of this assembly is approximately 1.5 K.

D. Interconnecting Coaxial Lines

The three coaxial lines, which interconnect the cryogenic input transmission line, cryogenic bandpass filter, cryogenic isolator, and HEMT amplifier input, are all 4-inch lengths of 0.141-inch semi-rigid copper coaxial cable. The total estimated noise contribution of these lines is 1.2 K, including connector losses and mismatch loss.

E. HEMT Amplifier Module

The three-stage HEMT/FET/FET amplifier module is shown in Fig. 3, and the schematic diagram of the unit is shown in Fig. 4. The input network, which can transform a 50-ohm source (Z0) to the desired source impedance, consists of a movable quarter-wave transmission line with a characteristic impedance of 35 ohms in cascade with an adjustable length of transmission line that has a characteristic impedance of 50 ohms. The real part of the source impedance is varied by changing the diameter and/or the surrounding dielectric of the sliding quarter-wave slug; the distance, L, between this slug and the HEMT device determines the reactance (and hence the resonant frequency) of the input network. A center conductor (0.072-inch diameter) in a square outer conductor (0.157-inch width), as suggested by Tomassetti [3], provides the 50-ohm transmission line (Z0). For the case of this particular amplifier, the quarter-wave sliding transformer (T1) is simply a rectangular Teflon slug that fills the coaxial line and is 0.915 inch in length. It has a 0.073-inch-diameter center hole which slides over the center conductor of the transmission line. Following the coaxial input matching circuit is a three-stage microstrip circuit board modified from a Berkshire Technologies commercial design.

The first-stage HEMT (Q1) is a 1/4-micrometer device supplied by General Electric. The second-stage MGF 1412 (Q2) and third-stage MGF 1402 (Q3) FET devices are supplied by Mitsubishi Electric. The source inductance feedback provided by L3 causes the optimum input VSWR of the module to coincide with minimum noise temperature. However, the bandwidth of the resulting input impedance match is narrow (approximately 100 MHz). A cryogenic isolator was included to improve input VSWR further and to ensure stability in the antenna environment. The source inductance feedback technique was suggested by Nevin and Wong [4] and was described for a cryogenically cooled GaAs FET amplifier by Williams, Lum, and Weinreb [5]. All inductors are made with 0.008-inch-diameter phosphor-bronze wire except for L3, which is made by the use of the source tab of the packaged HEMT together with a small grounding strap. Ferrite beads are used on the Q1 drain tab and on the Q2 gate and drain tabs to help quench parasitic oscillations at high gigahertz frequencies.

Zener diodes CR1 through CR6 (1N5232B) are employed on the gates and drains of the HEMT and FETs to protect them from overvoltage. An output 10-dB chip attenuator P1 is used to ensure >20 dB output return loss. Low-temperature HEMT performance is somewhat complicated by the formation of deep electron traps, which leads to "collapse" of the drain current–voltage characteristics at cryogenic temperatures [6]. This problem is alleviated by illuminating the HEMT junction with a low level of light. An LED mounted in the cover of the amplifier provides this function. The noise contribution (at 2295 MHz) of the completed three-stage amplifier is 4.6 K with 29 dB of gain (which includes the 10-dB output attenuator). The noise temperature of the amplifier module (see Fig. 5) was measured by using a Hewlett-Packard

HP 346B calibrated noise source connected to the amplifier input through a cooled 20-dB Narda Model 4779 attenuator.[1]

F. Closed-Cycle Refrigerator

The Model 350 Cryodyne refrigerator and compressor is supplied by Cryogenics Technology, Inc. (CTI). The refrigerator is enclosed in a vacuum housing and radiation shield designed specifically for DSN applications. The first stage of the refrigerator operates at 50 K and the second stage at 12 K. The refrigerator is designed to provide 15 watts of cooling capacity at 70 K and 3 watts of cooling capacity at 15 K. Cool-down time of the HEMT/CCR system is approximately 3 hours.

IV. DSS-13 Installation

The measured input noise temperature of the assembled HEMT/CCR package was 7.4 to 9.5 K across the 2200- to 2300-MHz range, as shown in Fig. 5. On August 19, 1986, the HEMT/CCR was first installed without the cryogenic bandpass filter. System noise temperature was measured with the antenna at zenith during clear weather. On August 22, 1986, the cryogenic bandpass filter was installed in the HEMT/CCR system, and the system noise temperature was again measured. The results of these measurements are shown in Fig. 6.

Since the time of installation, there have been no electronics failures. One drive unit failure occurred, which was attributed to oil carryover from a faulty compressor. The system has been in nearly continuous operation up to the time of this writing.

V. Conclusions

The 2.3-GHz HEMT/CCR system has proven, as expected, to be a more reliable system than the maser/CCR system. The noise temperature of future HEMT devices is expected to improve, and it is clear that the noise contribution of other cryogenic input components, such as the isolator and filter, can also be lowered. Therefore, the noise performance of this system can probably be made to surpass that of the DSN Block III TWM/CCR (5- to 8-K noise temperature) during the coming year.

[1] S. Weinreb and R. Harris, *Low Noise, 15 GHz, Cooled, GaAs FET Amplifier*, NRAO Internal Report No. 235, National Radio Astronomy Observatory, Charlottesville, Virginia, September 1983.

Acknowledgments

The author would like to extend special thanks to Jan Loreman, who designed and supplied the refrigerator and input transmission line, and to Juan Garnica and Chuck Goodson for the DSS-13 installation and for testing of the system.

References

[1] S. M. Petty, "Microwave Devices," in *Low-Temperature Electronics*, R. K. Kirschman (ed.), New York: IEEE Press, pp. 358–363, 1986.

[2] D. Norris, "Low-Noise Cryogenic Transmission Line," *TDA Progress Report 42-91*, vol. July–September 1987, Jet Propulsion Laboratory, Pasadena, California, November 15, 1987.

[3] G. Tomassetti, S. Weinreb, and K. Wellington, "Low Noise 10.7 GHz Cooled GaAs FET Amplifier," *Electronics Letter*, vol. 17, no. 25/26, pp. 949–951, December 10, 1981.

[4] L. Nevin and R. Wong, "L-Band GaAs FET Amplifier," *Microwave Journal*, vol. 22, no. 4, pp. 82–83 and 92, April 1979.

[5] D. Williams, W. Lum, and S. Weinreb, "L-Band Cryogenically Cooled GaAs FET Amplifier," *Microwave Journal*, pp. 73–76, October 1980.

[6] A. W. Swanson, "Pseudomorphic HEMT," *Microwaves and RF*, vol. 26, no. 3, pp. 139–150, March 1987.

Table 1. Noise temperature budget for 2.3-GHz HEMT/CCR system at 2295 MHz

System component	Noise contribution, K	
	With maser	With HEMT
Sky	3.5	3.5
Antenna	5.0	5.0
Horn	1.2	1.2
Combiner and yoke	10.0	10.0
54-dB coupler	0.2	0.2
Waveguide switch	0.5	0.5
Waveguide and bends	2.0	2.0
Waveguide bandpass filter	3.4	–
Dual cross guide coupler	0.4	0.4
Low noise amplifier	2.5	8.7
Total system noise temperature	28.7	31.5

Table 2. Estimated noise contributions to DSS-13 system noise temperature at 2295 MHz using maser/CCR and HEMT/CCR

System component	Noise contribution, K
Input line	0.3
Filter	0.9
Isolator	1.5
Coax	1.2
HEMT amplifier module	4.6
Follow-up amplifier	0.2
Total input noise temperature	8.7

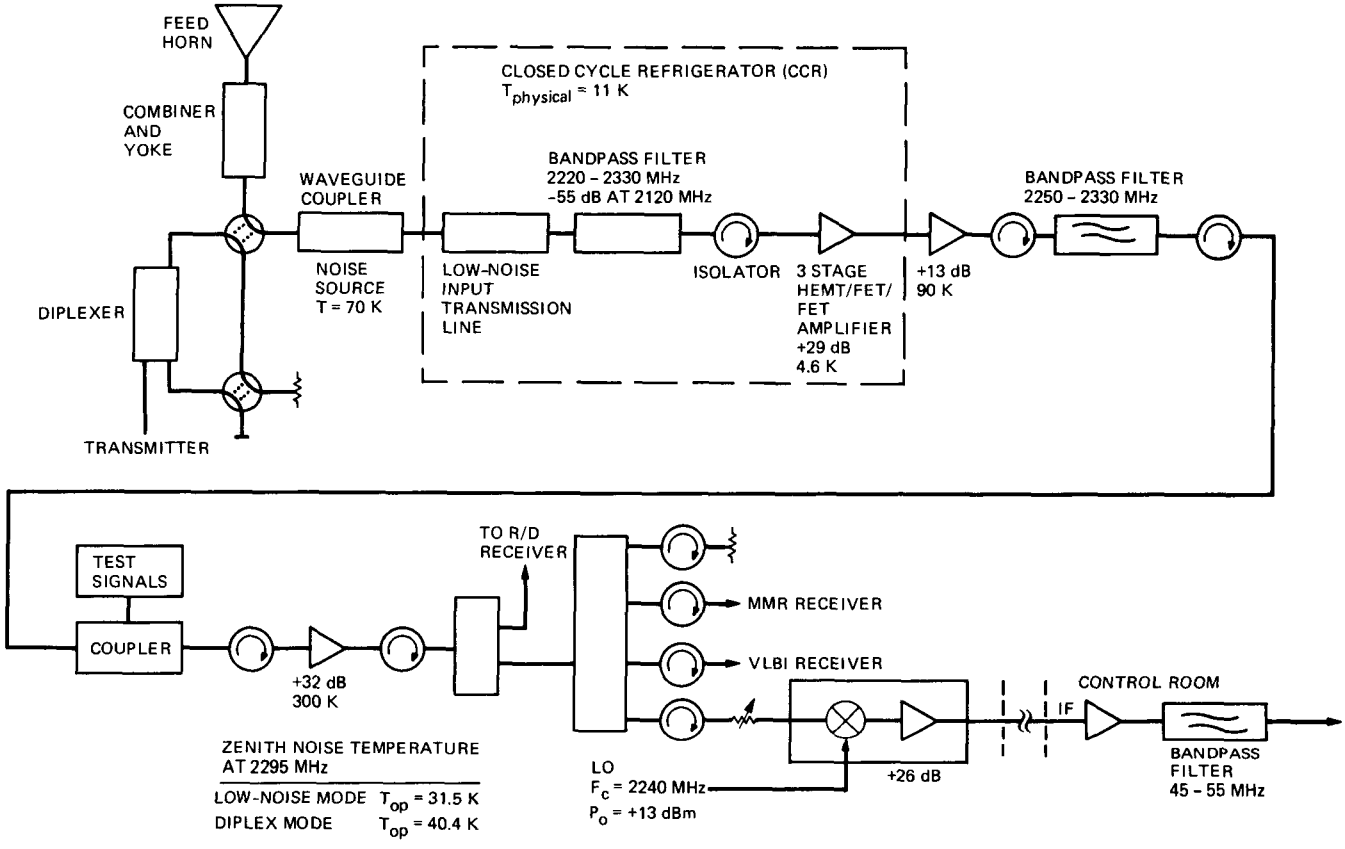


Fig. 1. DSS-13 S-band receiver system block diagram

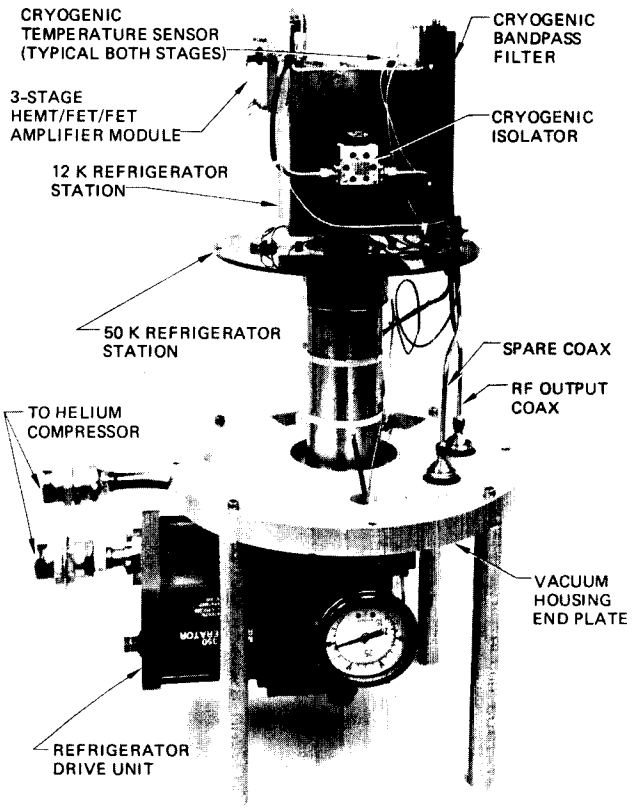


Fig. 2. HEMT/CCR package

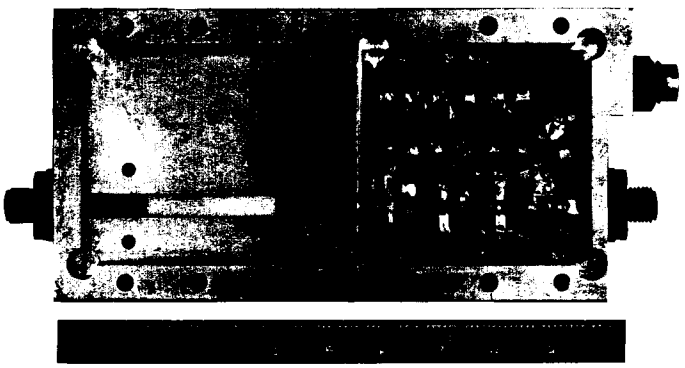


Fig. 3. HEMT/FET/FET amplifier module

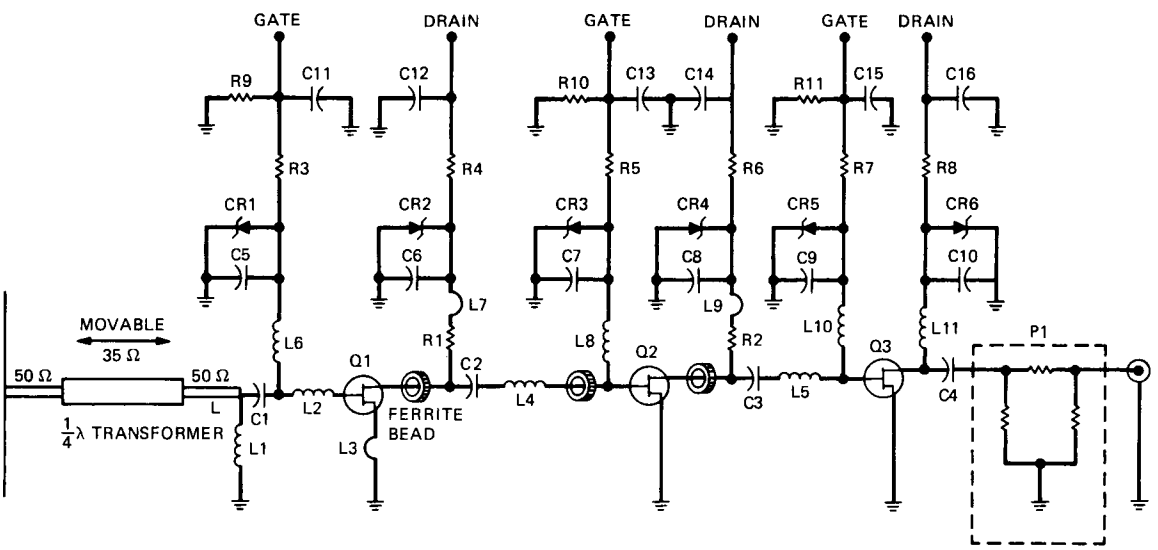


Fig. 4. Three-stage amplifier schematic diagram

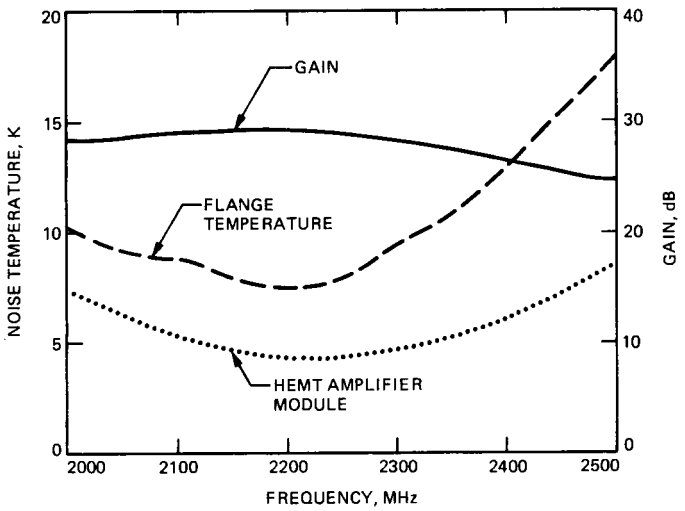


Fig. 5. Noise temperature of HEMT amplifier module and noise temperature (and gain) reference at the WR 430 waveguide flange with cryogenic filter not installed

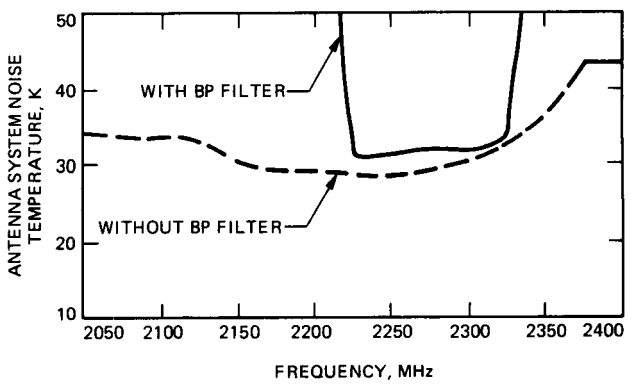


Fig. 6. DSS-13 antenna system noise temperature with and without the cryogenic bandpass filter

A Cold Ejector for Closed-Cycle Helium Refrigerators

D. L. Johnson
Radio Frequency and Microwave Subsystems Section

D. L. Daggett
San Jose State University, California

This article presents the test results of an initial cold helium ejector design that can be installed on a closed-cycle refrigerator to provide refrigeration at temperatures below 4.2 K. The ejector, test apparatus, instrumentation, and test results are described. Tests were conducted both at room temperature and at cryogenic temperatures to provide operational experience with the ejector as well as for future use in the subsequent design of an ejector that will provide refrigeration at temperatures below 3 K.

I. Introduction

Ejectors have been used for many years near ambient temperature in refrigeration equipment, boilers, and chemical processing equipment. The principle of the ejector lies in the ability of a moving jet of primary fluid to entrain a secondary fluid and move it downstream. The ejector, then, is a simplified type of jet pump or compressor which pulls in the low-pressure stream and increases the pressure of that stream by mixing it with the high-pressure stream.

Rietdijk [1] first proposed the use of an ejector in cryogenic refrigerators. Through use of a cold ejector in place of a Joule–Thomson expansion valve in a closed-cycle helium refrigerator, subatmospheric pressure can be created in the volume over a liquid helium bath. The ejector pumps on the vapor over the bath, reducing the vapor pressure over the bath and thereby lowering the temperature of the bath. Since this is accomplished at cryogenic temperatures, low-pressure-drop (large-diameter) tubing and heat exchangers are not required to maintain the low pressures in the gas line between the bath and an external vacuum pump at ambient temperatures. Thus, a lower temperature can be obtained without necessitating the use of a vacuum pump or a room-temperature compressor at subatmospheric pressure. This eliminates air leakage, saves power, and permits the use of smaller heat exchangers within the system.

Other benefits of the ejector are that it is a simple, lightweight mechanical device that has no moving parts, does not require additional compressor power, and will not introduce any additional contaminants. By replacing the Joule–Thomson expansion valve with an ejector and using proper thermodynamic optimization, a potential increase in the overall system efficiency of a 4.2-K closed-cycle refrigerator can be realized [1], [2].

One of the authors (D. L. Daggett) recently received a 10-week NASA Summer Fellowship at JPL to work on a cold helium ejector for a 4.5-K closed-cycle helium refrigerator. His task was to perform a one-dimensional thermodynamic analysis for a cold helium ejector and to design a cold helium ejec-

tor which, when installed in a 2-watt, 4.5-K closed-cycle refrigerator (CCR), would produce 0.5 watt of refrigeration at 2.5 K. Another objective was to determine if the ejector circuitry could be retrofitted into an existing CCR by installing it in place of the Joule-Thomson valve, leaving all other components in the CCR intact.

The analysis had to take into consideration the geometry of the primary converging-diverging nozzle, the diameter of the nozzle throat, the length and angle of divergence of the nozzle diffuser, the length and diameter of the mixing chamber, and the length and angle of divergence of the ejector diffuser section. Any ejector design resulting from the analysis would likely take weeks to fabricate, as the small size of the nozzle presents special problems in fabrication.

In efforts to gain some early experimental experience with an ejector to aid in the design analysis, an ejector of approximate size and geometry was formulated from the refrigeration characteristics of the 2-watt CCR and from an averaging of the ejector designs discussed in the literature [3]-[8]. This ejector was fabricated and installed in the CCR within the 10-week NASA Summer Fellowship period to provide some test results to aid in the analysis.

The results of the initial ejector-CCR test are included in this introductory report. A forthcoming detailed report will describe at length the analysis of the cold ejector, the ejector design, and the ejector-CCR test results.

II. Ejector Description

A schematic drawing of the ejector is shown in Fig. 1. The primary nozzle is a converging-diverging nozzle that acts to restrict the flow of gas in much the same way as the Joule-Thomson (J-T) valve in the CCR. The static high-pressure gas enters the ejector at the primary nozzle. As the gas moves through the converging portion of the nozzle, it is accelerated, converting the potential energy of the gas into kinetic energy. If the nozzle is properly designed for the gas flow, the gas reaches sonic velocity at the nozzle throat and continues to accelerate as the gas exits the nozzle. This high velocity jet is directed into a duct (called a mixing chamber) and induces a low-pressure secondary stream into the mixing chamber. The two streams exchange momentum in the mixing chamber and are discharged from the ejector at the end of the subsonic diffuser at an intermediate pressure.

The prototype ejector consisted of a non-adjustable primary converging-diverging nozzle assembly which was attached to a main ejector body (see Fig. 2). Precise concentric alignment of the primary nozzle with the mixing chamber is important in obtaining optimum performance. Thus, a positioning ring was installed around the primary nozzle to ensure concentric alignment with the ejector body. The nozzle was attached to the ejector by means of a stainless steel tube. This allowed some radial movement within the ejector body for the alignment ring to position the nozzle correctly. No attempt was made in this initial design to make the nozzle axial position an adjustable parameter. Rather, the fixed axial position of the nozzle with respect to the mixing chamber inlet was determined from calculations showing the required annular gap between the nozzle and the ejector body to produce a desired secondary flow rate. Differential thermal contractions in the ejector components had to be considered in the proper positioning of the nozzle for 4-K operation.

The ejector body and the nozzle were manufactured from brass. The primary and secondary inlets as well as the outlet area were machined with openings to allow stainless steel tubing to be inserted directly into the body and silver soldered in place. These tubes were used to plumb the ejector into the J-T circuit of the refrigerator. The two body parts of the ejector were soldered together with a low-temperature lead/tin solder for easy assembly and disassembly.

Manufacture of the components required precise machining as a result of the small dimensions and tight tolerances of the primary nozzle throat. Electro-discharge machining (EDM) was selected as the best way to fabricate all of the ejector components. A specially shaped electrode was manufactured to machine the converging portion of the primary nozzle. EDM was also used to form the 0.006-inch-diameter nozzle throat, the 0.030-inch-diameter mixing chamber, and the diffuser cones for both the nozzle and the ejector body.

III. Refrigerator Description

Figure 3 shows a comparison between the closed-cycle refrigerator with a conventional Joule-Thomson circuit and the modified refrigerator using a cold ejector circuit. In the normal J-T cycle, compressed gas is cooled to below its inversion temperature and expanded through a small orifice. The J-T effect makes use of an isenthalpic expansion of the gas, transforming the high-level static pressure of the gas into kinetic energy as it passes through the small expansion valve orifice. Upon expansion, all of the kinetic energy is dissipated. The net effect is a decrease in the temperature of the fluid (or an increased condensation of the liquid).

Rietdijk [1] showed that it was possible to use part of this otherwise wasted kinetic energy to compress an amount of gas from a secondary loop in the ejector circuit. The high-pressure gas enters the ejector at the primary nozzle (Fig. 1)

and is accelerated as it passes through the primary nozzle, causing the potential energy of the gas to be converted to kinetic energy. While the gas expands through the nozzle in a nearly isentropic fashion, a temperature decrease of the gas results. This high-velocity jet stream entrains a low-pressure secondary flow that enters the ejector at the suction port. The two streams exchange momentum in the mixing chamber and are discharged from the ejector at the end of the subsonic diffuser at an intermediate pressure. The discharge will exist as a two-phase fluid. The liquid portion is separated off and forms the secondary mass flow, passing through an optional heat exchanger and through a J–T expansion valve. The helium undergoes a slight isenthalpic expansion in the J–T expansion valve. The J–T valve is used to restrict the secondary mass flow. The liquid from this expansion is collected and used for refrigeration in a secondary pot. The low-temperature vapor from the evaporated liquid flows through the heat exchanger and into the suction side of the ejector to complete the cycle.

Note that in the J–T circuit of the conventional refrigerator, the temperature is dictated by the suction pressure of the compressor. In the ejector circuit, for the same pressure of the external compressor, a lower temperature can be achieved because of the additional pumping effect of the high velocity jet.

IV. Refrigerator Configuration

The closed-cycle refrigeration system used in the tests described below is shown in Fig. 4. Before modification for the ejector circuit, the CCR with a fixed-orifice J–T valve provided 2 watts of refrigeration at 4.5 K with a mass flow rate of 2.6 scfm through the J–T circuit. The CCR uses a two-stage Gifford–McMahon expansion engine (CTI Model 350) to provide an intermediate refrigeration of approximately 25 watts at 60 K and 5 watts at 15 K. The heat exchangers and the J–T valve were built at JPL. The CCR uses a 5-hp Dunham–Bush compressor to provide 30 cfm of helium gas to the CCR. A detailed description of this refrigerator is presented in [9].

To install the cold ejector circuit, the 4-K station cold plate on which the maser is normally mounted was removed. In its place, a larger 4-K storage pot was installed which allowed ample volume for the phase separation of the fluid. The fixed-orifice J–T valve was removed, and the ejector was installed in its place. An adjustable J–T valve was installed to provide flow restriction in the secondary loop. The J–T adjustment is made using a Starret micrometer that could be adjusted from outside the cryostat. The micrometer was connected by a long, slender stainless steel rod to a finely tapered needle in the J–T valve body. A 10-cc volume copper pot was installed as the sub-4-K station in the secondary loop.

The ejector test measurements included static pressures, temperature, and mass flow at locations shown in Fig. 5. The pressures were measured at ambient temperature using Endevco Model 8530A pressure transducers connected by 0.063-inch-OD capillary tubing to the locations of interest. The intermediate temperatures of the CCR's expansion engine were measured using Lake Shore DT-500 silicon diodes. The low temperatures of the ejector circuit were measured using calibrated carbon glass resistors, also from Lake Shore Cryotronics. Small heater resistors were placed on the 4-K station and the secondary station to measure the refrigeration capacity of each station. A heater was also attached to the ejector inlet stream to permit variation of the temperature of the high pressure primary gas stream. The primary mass flow rate of the ejector circuit was measured on the return stream to the compressor at ambient temperature with a Hastings Linear Mass Flowmeter, Model NAHL-5, calibrated for helium. The flow rate for the secondary gas stream was measured during room temperature testing through the insertion of a small flowmeter in the secondary gas loop. When operating at liquid helium temperatures, the secondary gas stream could not be monitored; the cold secondary flow rate, however, can be estimated given knowledge of the temperature, pressure, and refrigeration capacity of the secondary loop.

V. Results

The ejector was designed to operate with a primary pressure of 300 psia and a primary mass flow rate of 2.5 scfm, consistent with the operation of the 2-watt CCR with the J–T valve. The compressor return pressure was set to the same level for both configurations such that the ejector exit stream pressure would be 16–17 psia (operating temperature of 4.4 K). Unfortunately, the nozzle diameter was machined larger (0.006 inch) than was specified (0.005 inch), such that the mass flow rate through the nozzle at ambient temperatures was on the order of 0.6 scfm. This large flow rate was a clear indication that at 4 K the gas flow through the nozzle for the 300-psia operating pressure would be greater than that which the expansion engine and the heat exchangers in the circuit would be able to cool effectively. In fact, it would also tax the 5-hp helium compressor.

The low-temperature refrigeration capacity of the secondary loop of the ejector depends on the helium mass flow rate and on the vapor pressure of the secondary stream. A high flow rate yields a high capacity (heat of vaporization of the liquid), and the vapor pressure will determine the operating temperature of the secondary loop. Thus, the ejector operation requires a high rate of entrainment of the secondary stream by the primary stream to provide a large refrigeration capacity and to ensure a low secondary vapor pressure. Therefore, parameters of importance include the pressure P_1, the tem-

perature T_1, and the mass flow rate $\dot{m}_1$ of the primary flow stream; the suction pressure P_2 and the mass flow rate $\dot{m}_2$ of the secondary stream; and the ejector discharge pressure P_3. For interpretation purposes, the data is often plotted as the ratio of the ejector discharge pressure to the secondary pressure P_3/P_2, and as the mass entrainment ratio $\dot{m}_2/\dot{m}_1$ (secondary mass flow rate over primary mass flow rate). Both P_2 and $\dot{m}_2$ can be varied by adjusting the position of the primary nozzle relative to the mixing chamber inlet and by adjusting the restriction of the J-T valve. In these tests, however, the nozzle position could not be adjusted; its fixed position was determined from calculations prior to assembly of the ejector.

The only two parameters that could be adjusted in the tests were the primary pressure P_1 and the J-T valve restriction. The primary pressure was varied from atmospheric pressure (14.5 psia) for calibrating the pressure sensors to a high-pressure 310 psia. The high pressure is representative of the upper pressure limit of an operating helium compressor for the CCR. The primary mass flow rate through the nozzle was linear with respect to the primary pressure; the proportionality constant is dependent on the nozzle throat diameter. At 312 psia, the primary mass flow rate was 0.55 scfm for the room temperature measurements. The J-T restriction in the secondary loop was controlled with micrometer adjustment of the J-T valve. The micrometer could be adjusted from a fully restricted state (a 0.108 setting) to fully open (a 0.400 or larger setting).

Figure 6 shows the relationship between the mass flow rate and the suction pressure for the secondary stream as a function of the primary stream flow rate. For a given primary flow rate, the position along the curve gives an indication of the J-T restriction, with a fully restricted valve represented as the lower left end of the curve. From this figure, it is clear that a high primary stream flow rate is required if both a high secondary flow rate and a low suction pressure are desired.

Figure 7 shows the relationship between the pressure ratio and the mass entrainment ratio for several primary mass flow rates. It may be observed from the figure that there is a practical limit in the size of the primary that will optimize both the suction pressure and the entrainment ratio. For low entrainment ratios, the pressure ratio increases with increasing primary flow rates. As the entrainment ratio increases, however, this effect becomes less and less pronounced until finally this relation is reversed, with the pressure ratio decreasing with increasing primary flow rates. This implies that for a given secondary mass flow rate, there is a maximum pressure ratio that may be achieved by varying primary flow rates. To what degree this room temperature relationship also holds at cryogenic temperatures is uncertain, as it was not possible to vary the parameters sufficiently at cryogenic temperatures to measure this effect. However, it is understood that the efficiency of the nozzle and ejector design will have a large effect on the quality of the data.

Sample data sets for the ejector operating at cryogenic temperatures are shown in Table 1. The large flow rate of the primary stream made it difficult to vary the primary pressure, the heat loads, and the J-T micrometer positions (flow restrictions) to any great degree. The primary stream pressure could not be increased to much above 230 psia without warming the ejector. Likewise, the J-T valve could just barely be opened. Opening the J-T micrometer above 0.150 brought the temperature difference between T_2 and T_3 to near zero. The pressure ratios as a function of the primary pressure for the cold data are very consistent with the pressure ratios for the room temperature data. These data have been plotted in Fig. 8, along with the room temperature data, for two J-T flow restrictions. This correlation may prove useful in providing a dimensional analysis by which an ejector's cryogenic performance may be predicted by room temperature tests.

VI. Conclusions

The test results of this first ejector design have been very helpful in furthering an understanding of the ejector operation at ambient and cryogenic temperatures. The warm temperature tests clearly showed the ability of the primary jet to pump the secondary gas stream, with pressure ratios reaching 6.5 in this first design. Both the pressure ratio and the entrainment ratio are very dependent upon the primary stream mass flow rate and the J-T restriction. At large J-T restrictions, the pressure ratio of the streams at cryogenic temperatures was quite consistent with the room temperature pressure ratios. This fact could aid in the performance of preliminary room temperature tests of an ejector design to anticipate cryogenic temperature performance.

These initial test results have also been of benefit in the ongoing thermodynamic analysis of an "optimized" ejector design. The flow rate through the ejector nozzle was larger than predicted, and thus the new ejector nozzle throat will be made slightly smaller than the calculations indicate. A new ejector nozzle having a 0.004-inch-diameter throat is currently being fabricated.

References

[1] J. A. Rietdijk, "The Expansion-Ejector: A New Device for Liquefaction and Refrigeration at 4K and Lower," in *Liquid Helium Technology, Proceedings of the International Institute of Refrigeration*, New York: Pergamon Press, 1966.

[2] F. W. Pirtle, P. A. Lessard, J. M. Kaufman, and P. J. Kerney, "Thermodynamic Aspects of Small 4.2-K Cryocoolers," in *Advances in Cryogenic Engineering*, vol. 27, pp. 595–602, New York: Plenum Press, 1982.

[3] F. W. Pirtle and P. J. Kerney, "Josephson Junction Cryocooler System Design and Component Development Program," Technical Report AFWAL-TR-84-3087, Flight Dynamics Laboratory, Wright-Patterson Air Force Base, Ohio, February 1985.

[4] A. I. Ageev, N. N. Agapov, and V. A. Belushkin, "Experimental Study on Helium Cryogenic Ejector" (in Russian), No. 8-8608, presented at the Joint Institute for Nuclear Research, Dubna, USSR, 1975.

[5] V. P. Beliakov, V. I. Epifanova, V. S. Baikov, T. M. Rosenoer, and V. V. Usanov, "Helium Microejectors," in *Proceedings of the Thirteenth International Congress of Refrigeration*, p. 65, 1971.

[6] P. Bernheim and E. Legrives, "Les Éjecteurs Cryogéniques Helium," presented at the IRF–IIR Commission AI/2 Conference, Paris, 1983.

[7] P. Rudolf von Rohr and C. Trepp, "Experimental Investigation of an Ejector," *Cryogenics*, vol. 25, pp. 684–686, 1985.

[8] P. Rudolf von Rohr, "Tieftemperaturanlagen für die Kühlung supraleitender Systeme," doctoral dissertation, Institut für Verfahrens und Kältetechnik der ETH, Zürich, Switzerland, 1983.

[9] M. Britcliffe, "A 2-Watt, 4-Kelvin Closed-Cycle Refrigerator," *TDA Progress Report 42-91*, vol. July–September 1987, Jet Propulsion Laboratory, Pasadena, California, November 15, 1987.

Table 1. Cold ejector performance data

P_1, psia	T_1, K	P_2, psia	T_2, K	P_3, psia	T_3, K	$\dot{m}_1$, scfm	$\dot{Q}_2$, mW	J-T setting	P_3/P_2
177	7.1	12.7	4.09	22.2	4.68	4.51	0	0.125	1.75
182	7.7	12.1	4.03	21.8	4.65	4.23	0	0.115	1.80
201	7.7	10.9	3.95	22.5	4.70	4.66	0	0.115	2.07
228	7.7	9.33	3.79	23.3	4.75	5.10	0	0.115	2.50
231	7.8	9.47	3.84	23.7	4.75	5.19	0	0.115	2.50
201	7.7	11.0	3.97	22.4	4.68	4.64	200	0.115	2.04
202	7.5	13.0	4.12	22.5	4.70	4.70	100	0.150	1.73

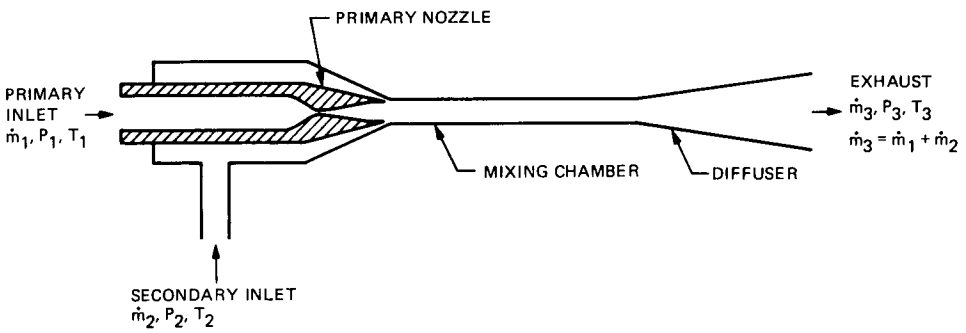


Fig. 1. Schematic of ejector configuration

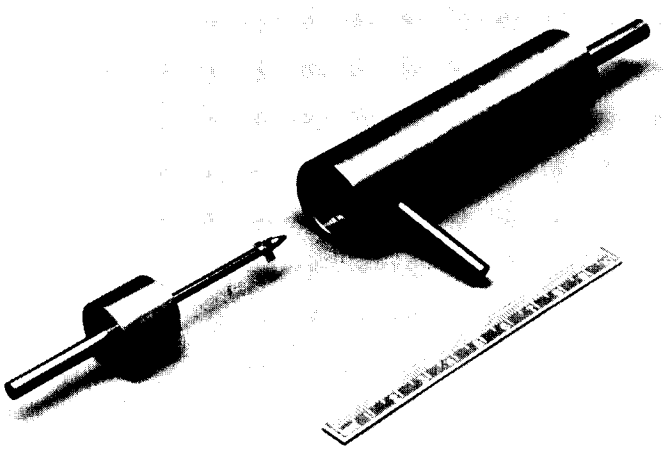


Fig. 2. Expanded view of cold helium ejector

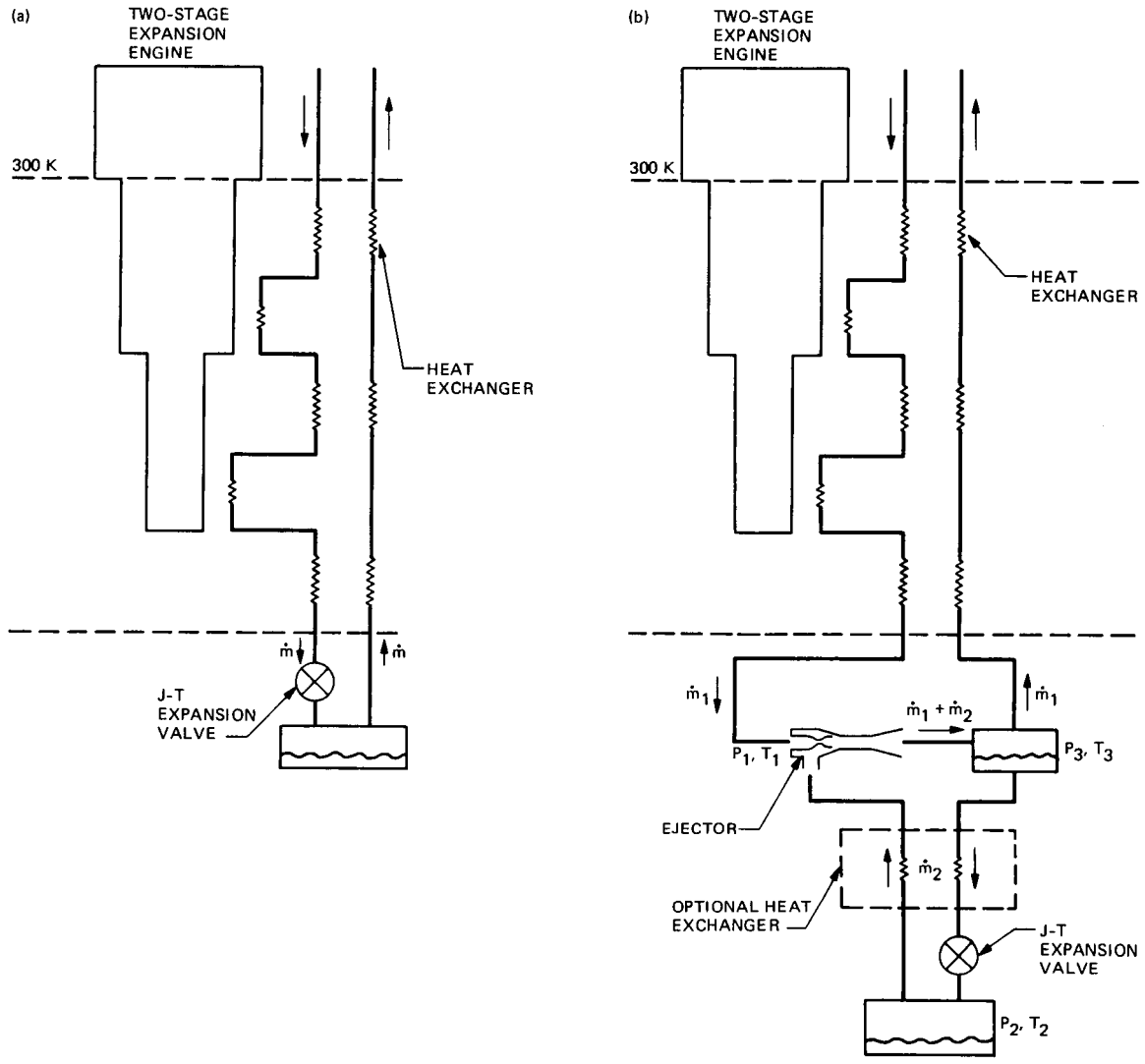


Fig. 3. Refrigeration circuits for (a) the conventional Joule-Thomson cycle and (b) the Joule-Thomson cycle as modified with cold ejector

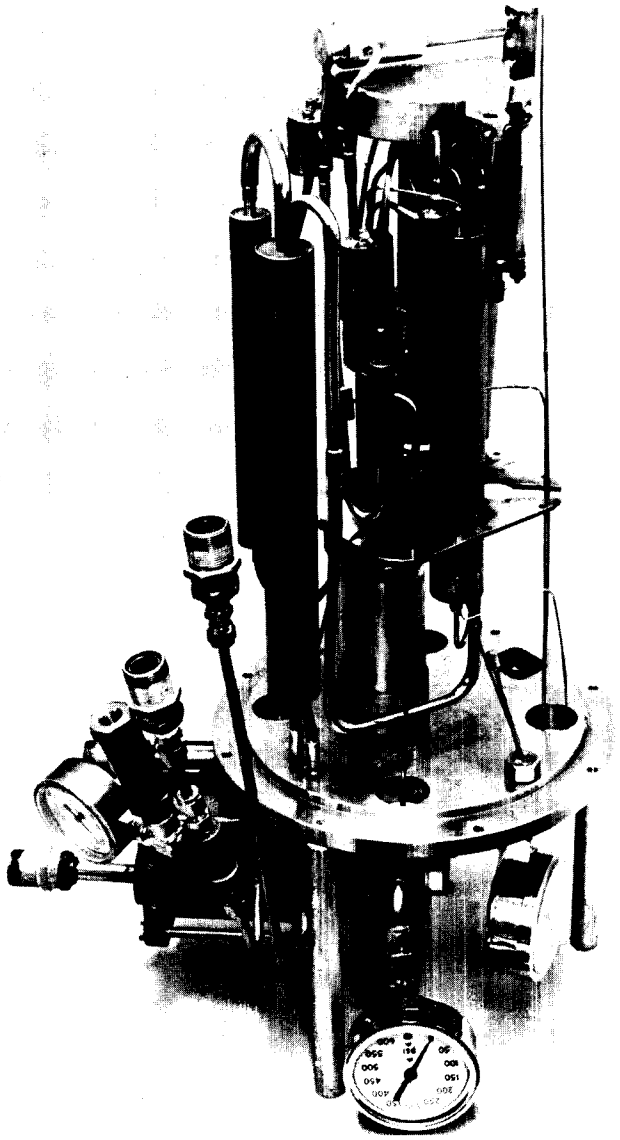


Fig. 4. Closed-cycle helium refrigerator modified with cold ejector for laboratory testing

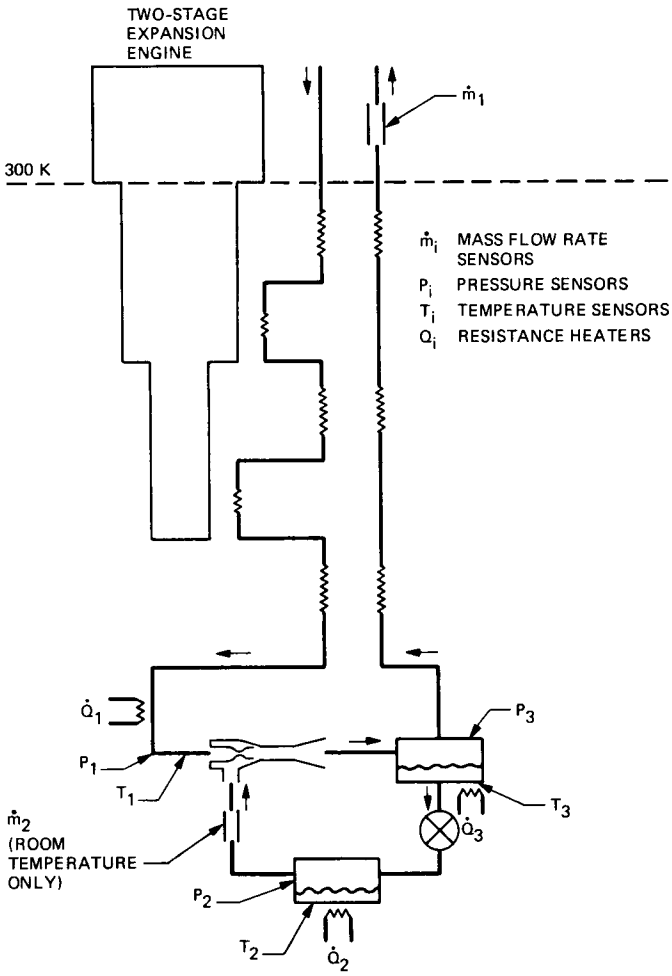


Fig. 5. Refrigerator-ejector schematic showing location of sensors

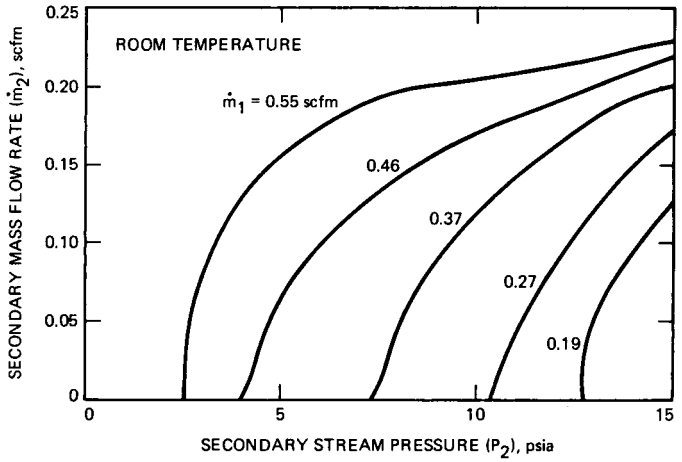


Fig. 6. Secondary mass flow rate as a function of the secondary stream suction pressure for different primary mass flow rates

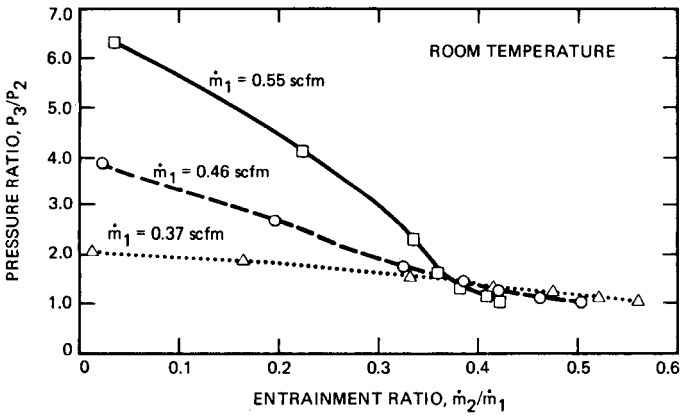


Fig. 7. Pressure ratio as a function of the entrainment ratio for different primary mass flow rates

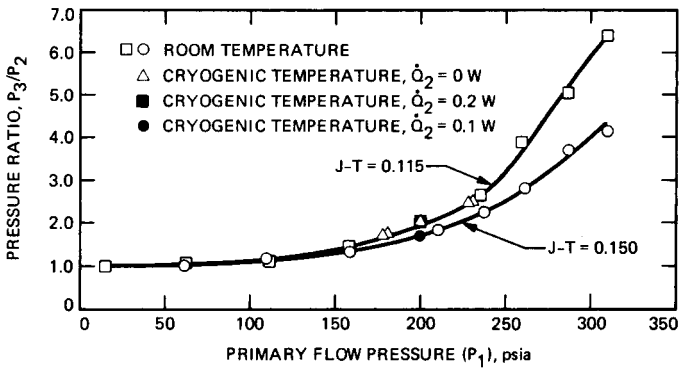


Fig. 8. Comparison of room temperature and cryogenic temperature pressure ratios as a function of primary stream pressure for two different J-T flow restrictions

Frequency Doubling Conversion Efficiencies for Deep Space Optical Communications

D. L. Robinson and R. L. Shelton
Communications Systems Research Section

The theory of optical frequency doubling conversion efficiency is analyzed for the small signal input case along with the strong signal depleted input case. Angle phase matching and beam focus spot size are discussed and design trades are described which maximize conversion efficiency. Experimental conversion efficiencies from the literature, which are less than theoretical results at higher input intensities due to saturation, reconversion, and higher-order processes, are applied to a case study of an optical communications link from Saturn. Double pass conversion efficiencies as high as 45 percent are expected. It is believed that even higher conversion efficiencies can be obtained using multipass conversion.

I. Introduction

In a deep space optical communications link, a frequency doubled Nd:YAG laser transmitter emitting at 0.532 μm is the preferred transmitter. When the Nd:YAG laser is diode pumped the laser is completely solid state and has higher reliability for spacecraft communication than its gas and dye counterparts. The Nd:YAG laser fundamentally outputs at 1.06 μm. Since detectors typically have very low quantum efficiencies at 1.06 μm, the laser is frequency doubled to 0.532 μm to produce a wavelength that has higher detector quantum efficiencies. In this article, frequency doubling conversion efficiencies are discussed. First, in Section II, theoretical conversion efficiencies are discussed. Then, in Section III, experimental results from the literature are compared to the theoretical results presented in Section II. A case study is then presented that uses experimental data from the literature and typical optical communication link parameters for a link from Saturn. Finally, the conclusion is given in Section IV.

II. Theoretical Analysis

In 1961, second harmonic generation was first experimentally demonstrated with a conversion efficiency of approximately 10^{-8} [1]. Today conversion efficiencies of 30 to 40 percent are not uncommon. Second harmonic generation conversion efficiency for plane waves can be expressed as:

$$\eta = \frac{P(2\omega)}{P(\omega)} = 2\left(\frac{\mu_o}{\epsilon_o}\right)^{3/2} \frac{\omega^2 (d_{eff})^2 L^2}{n^3} \left(\frac{P(\omega)}{\pi \omega_o^2}\right) \frac{\sin^2\left(\frac{\Delta kL}{2}\right)}{\left(\frac{\Delta kL}{2}\right)^2} \quad (1)$$

for small input intensities (see Appendix A [2]). Here, $P(2\omega)/P(\omega)$ is the conversion efficiency, η, d_{eff} is the effective nonlinear coefficient and ω is the fundamental frequency. The term $P(\omega)$ denotes the input power while $P(2\omega)$ denotes the second harmonic power. The variable L is the crystal length and Δk equals the difference between the wave vector ($k=2\pi/\lambda$), k_2, for the second harmonic frequency, 2ω, and the wave vector k_1 for the fundamental frequency, ω. The beam waist radius is ω_o and n is the index of refraction within the crystal. The quantities ϵ_o and μ_o are the electric and magnetic permeabilities, respectively, within a vacuum.

In Eq. (1), the $\sin^2 (\Delta kL/2)/(\Delta kL/2)^2$ factor can be made to approach unity, thus maximizing the conversion efficiency, with the proper phase matching techniques. Therefore, an understanding of phase matching is in order. If a second harmonic wave is produced at one plane and propagates to a second plane, and if another second harmonic wave is generated independently at this second plane, then the two waves interfere. The $\sin^2 (\Delta kL/2)/(\Delta kL/2)^2$ factor describes that interference, where $\Delta k = k_2 - 2k_1$. For maximum efficient frequency doubling, $\Delta k = 0$. Therefore, $k_2 = 2k_1$. A more detailed discussion of phase matching is given in Appendix B.

Looking at our initial conversion efficiency formula, Eq. (1), we have an inverse relationship between conversion efficiency and beam cross sectional area, $\omega_o^2 \pi$, assuming a uniform beam. This might lead one to tightly focus the beam. But because a tight focus results in greater divergence away from the beam waist, as Fig. 1 shows, minimizing the beam waist does not necessarily maximize the conversion efficiency. The confocal parameter, z, is defined as the distance from the plane of the beam waist to the plane in which the beam's cross sectional area is twice that at the beam waist (see Fig. 1).

$$z = \frac{\pi n \omega_o^2}{\lambda} \tag{2}$$

It should be noted that some authors define the confocal parameter as twice that distance [3].

If the confocal parameter is much larger than the crystal length, then the beam's cross sectional area is relatively constant through the crystal. The plane wave formulation approximates this case. However, if the crystal length is much greater than $2z$, the decreased intensity away from the beam waist results in a lower overall conversion efficiency and a focused wave formulation must be used.

If $L=2z$, we have confocal focusing. By using $L=2z=2\pi n\omega_o^2/\lambda$, area $= \pi\omega_o^2$, and $\lambda = 2\pi c/\omega$ we obtain an equation for frequency doubling conversion efficiency for the focused case:

$$\frac{P(2\omega)}{P(\omega)}_{\text{confocal focusing}} = \frac{2}{\pi c}\left(\frac{\mu_o}{\epsilon_o}\right)^{3/2} \frac{\omega^3 d^2 L}{n^2} \times P(\omega) \frac{\sin^2\left(\frac{\Delta kL}{2}\right)}{\left(\frac{\Delta kL}{2}\right)^2} \tag{3}$$

This formulation exhibits a conversion efficiency proportional to the crystal length, whereas in the plane wave model the conversion efficiency is proportional to L^2.

An exact analysis by Boyd and Kleinman [4] found that the conversion efficiency versus beam waist relationship follows the curve shown in Fig. 2. The optimum conversion results when $L = 5.68z$. This corresponds to an efficiency 20 percent greater than if $L = 2z$. Boyd and Kleinman's analysis incorporated an efficiency reduction factor in the conversion efficiency formula to analytically include the effect of focusing within the crystal. The efficiency reduction factor, h, is a function of crystal length, L, and the beam's confocal parameter, z. When $h(L/2z)$ enters the small input formula, Eq. (1), the yield is [4]:

$$\eta = 2\left(\frac{\mu_o}{\epsilon_o}\right)^{3/2} \frac{\omega^2 d_{eff}^2 L^2}{n^3} \left(\frac{P(\omega)}{\pi \omega_o^2}\right) h(L/2z) \tag{4}$$

In Eq. 4 we assume complete phase matching ($\text{sinc}^2(\Delta kL/2) = 1$) and observe that the conversion efficiency is proportional to $h(L/2z)$. Therefore, conversion efficiency is maximized with the maximum h. Figure 2 illustrates the $h(L/2z)$ function versus $L/2z$. Here, and in Eq. (4), we have assumed the most efficient case where Boyd and Kleinman's double refraction parameter, which includes the effect of the walk-off angle between the Poynting vectors of the fundamental and the second harmonic waves [3], has been set to zero.

All the previous formulations required that conversion to the second harmonic be small. In other words, the amount of fundamental light traversing the crystal available for conversion to the second harmonic remained constant throughout the crystal. Since an optical communications application requires higher power levels with much greater conversion efficiencies, a discussion of the depleted input situation is warranted. In the depleted input case conversion efficiency is high; therefore, the amount of fundamental light traversing the crystal is continually reduced; hence, the term depleted input.

The equation for conversion efficiency for the depleted input case is given in Eq. (5) (see Appendix C):

$$\frac{P(2\omega)}{P(\omega)} = 2\tanh^2\left[d_{\text{eff}}\left(\frac{\mu_o}{\epsilon_o}\right)^{3/4}\frac{\omega z}{n^{3/2}}\sqrt{\frac{P(\omega)}{\pi\omega_o^2}}\right] \quad (5)$$

Single pass conversion efficiencies for the small signal input and depleted input formula are plotted in Fig. 3 for comparison. In Fig. 3, the upper curve is the small signal input formulation and the lower curve is the depleted input formulation for frequency doubling of 1064 nm radiation in a 5.1 mm KTP crystal. As can be seen from Fig. 3 the small signal formula is only valid for a small region, up to a conversion efficiency of approximately 15 percent. Beyond that region, the loss of input intensity due to conversion to the second harmonic is substantial and the small signal formula breaks down. The intensity of the beam decreases as the beam propagates through the crystal; thus, the conversion efficiency in the crystal decreases. Equation (5) assumes plane wave propagation within the crystal; however, the analysis by Boyd and Kleinman for focusing within the crystal could be applied to the depleted input case as well. Since an optical communication application requires high conversion efficiencies, the depleted input situation is appropriate.

III. Case Study

Let us examine a specific nonlinear frequency doubling crystal, $KTiOPO_4$ (KTP). The KTP crystal is a relatively new second harmonic generation material and has a few advantages over such predecessors as KDP and $LiNbO_3$. It has a high damage threshold and a high nonlinear coefficient, as well as excellent optical quality.

The d_{eff} values within the literature vary between theoretical and experimental results. By averaging the values of d_{eff}, which were determined from the graphical data presented in a Lockheed study of KTP [5], a value of 2.2×10^{-23} m/V was found. In Fig. 4 experimental values for single pass second-harmonic-generation conversion efficiency for a guassian beam have been graphed with theoretical values calculated in Section II for the depleted input case. At fundamental intensities less than 80 MW/cm^2 the two curves are in agreement. However, for intensities greater than 80 MW/cm^2 the experimental values are less than the expected theoretical results. Driscoll *et al.* attribute this saturation of reconversion to the fundamental and higher-order processes [5]. Driscoll *et al.* achieved higher conversion efficiencies with a multimode laser, probably due to localized higher intensity within the modal spot structure; however, for our application of deep space optical communication, we are primarily interested in single mode TEM_{oo} operation. Due to the difference in theory versus experiment, we use Driscoll's experimental data to calculate conversion efficiency in the following case study.

A specific case for frequency doubling is a cavity dumping scheme intended for an M-ary pulse position modulated (PPM) optical communications link [6]. As an example, assume a pulse repetition rate of 14.3 kHz (114 kbits/s for M=256) with a dead time of 44.4 μs and a pulse width of 20 ns. These parameters are typical of links being considered from Saturn to Earth. If an average laser power of 500 mW is assumed then the peak power is 1.75 kW with 35 μjoules as the corresponding pulse energy. An average spot diameter of 100 μm inside the crystal corresponds to a power density in the crystal of 22 MW/cm^2. Driscoll's graph of experimental data [5] for double pass efficiency through a 5.1 mm KTP crystal relates a 22 MW/cm^2 beam to approximately 28 percent double pass efficiency as shown in Fig. 5 with an arrow.

That conversion efficiency can be improved upon by decreasing the repetition rate, decreasing the average spot size, or increasing the average power level. Table 1 summarizes the results for various system parameters. Note that η in Table 1 only results from a double pass senario [5] and will probably be higher for multipass configurations.

A better understanding of the roles of spot size and pulse width can be gained by breaking down Table 1 and graphing the information. Figure 6 plots average power in watts versus double pass conversion efficiency for different spot sizes. Figure 6 shows that for the 50 μm beam, the conversion efficiency peaks and begins to drop off at approximately 500 mW, while the beam with a spot size of 100 μm peaks and begins to drop off at 2 watts. For two beams of the same average power, halving the spot diameter quadruples the intensity. At high intensities, the conversion efficiency decreases due to saturation, reconversion, higher order processes, and intensity fluctuations [5]. Thus, the curve for the 50 μm beam peaks and begins to dip at average power levels which are one fourth of those for similar efficiencies with a 100 μm beam. Figure 7 displays curves of double pass conversion efficiency versus average input power with a 10 ns pulse width and a 20 ns pulse width. For the specific case calculated, conversion efficiency peaks at 1.25 watts for the 10 ns pulse width case at a nominal pulse repetition rate of 14.3 kHz as discussed earlier. The 20 ns pulse width case peaks at a slightly higher average power level of 2.0 watts for the same repetition rate.

Furthermore, by varying the parameters, as the graphs and table display, double pass conversion efficiencies from 28 percent to 45 percent should be attainable for an optical communications link. However, even higher efficiencies are believed possible with a multipass system.

IV. Conclusion

A theoretical analysis of frequency doubling conversion efficiency has been presented along with a case study of an optical communications link from Saturn. The case study used experimental test results found in the literature and applied them to typical link parameters. An average power, 1 watt, laser operating at over 100 kbits/s in a pulse position modulation mode with M=256, and a pulse width of 10 ns, yielded double pass conversion efficiencies as high as 45 percent. Flux levels in this case, for a uniform average 100 μm diameter spot size, were approximately 90 MW/cm^2 which was well below the damage threshold of ~350 MW/cm^2. Experimental conversion efficiencies found in the literature at this intensity level were in close agreement with those calculated from the theoretical discussion presented in Section I. At higher intensity levels, experimental results were less than those calculated due to saturation from reconversion and higher-order processes. With a multipass configuration, it is possible that still higher conversion efficiencies may be achievable.

References

[1] P. A. Franken, A. E. Hill, C. W. Peters, and G. Weinreich, "Generation of Optical Harmonics," *Phys. Rev. Letters*, vol. 7, pp. 118–119, 1961.

[2] A. Yariv, *Quantum Electronics*, New York: John Wiley and Sons, Inc., 1975.

[3] Y. R. Shen, *The Principles of Nonlinear Optics*, New York: John Wiley and Sons, Inc., 1984.

[4] G. D. Boyd and D. A. Kleinman, "Parametric Interaction of Focused Gaussian Light Beams," *Journal of Applied Physics*, vol. 39, no. 8, pp. 3597–3639, July 1968.

[5] T. A. Driscoll, H. J. Hoffman, R. E. Stone, and P. E. Perkins, "Efficient Second-Harmonic Generation in KTP Crystals," *J. Opt. Soc. Am. B*, vol. 3, no. 5, pp. 683–686, May 1986.

[6] D. L. Robinson, "A Novel Approach to a PPM Modulated Frequency Doubled Electro-Optic Cavity Dumped Nd:YAG Laser," to be published in JPL TDA Progress Report.

Table 1. Summary of various system parameters

Spot diameter	Pulse width	Dead time	Rep rate M=256	Average power	Peak power	Power density	η
100 μm	20 ns	44.4 μs	14.3 kHz	500 mW	1.8 kW	22 MW/cm^2	28%
100 μm	20 ns	44.4 μs	14.3 kHz	1 W	3.5 kW	45 MW/cm^2	40%
100 μm	20 ns	44.4 μs	14.3 kHz	2 W	7.0 kW	89 MW/cm^2	45%
100 μm	20 ns	44.4 μs	14.3 kHz	3 W	10.5 kW	134 MW/cm^2	43%
50 μm	20 ns	44.4 μs	14.3 kHz	250 mW	0.88 kW	45 MW/cm^2	41%
50 μm	20 ns	44.4 μs	14.3 kHz	500 mW	1.8 kW	89 MW/cm^2	45%
50 μm	20 ns	44.4 μs	14.3 kHz	1 W	3.5 kW	179 MW/cm^2	43%
50 μm	20 ns	44.4 μs	14.3 kHz	2 W	7.0 kW	357 MW/cm^2*	40%
100 μm	10 ns	44.4 μs	14.3 kHz	500 mW	3.5 kW	45 MW/cm^2	40%
100 μm	10 ns	44.4 μs	14.3 kHz	1 W	7.0 kW	89 MW/cm^2	45%
100 μm	10 ns	44.4 μs	14.3 kHz	2 W	14.0 kW	178 MW/cm^2	43%

*damaged threshold

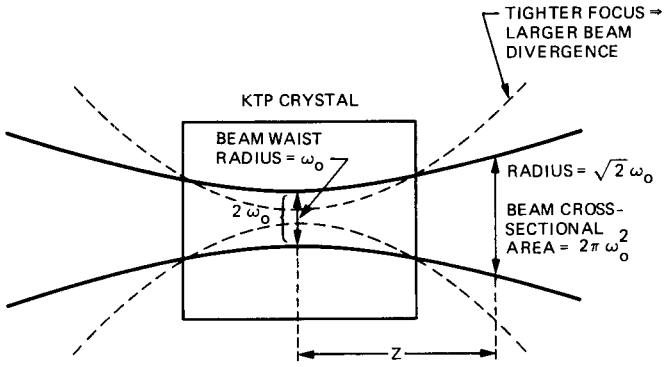


Fig. 1. Focusing of the beam through the frequency doubling crystal

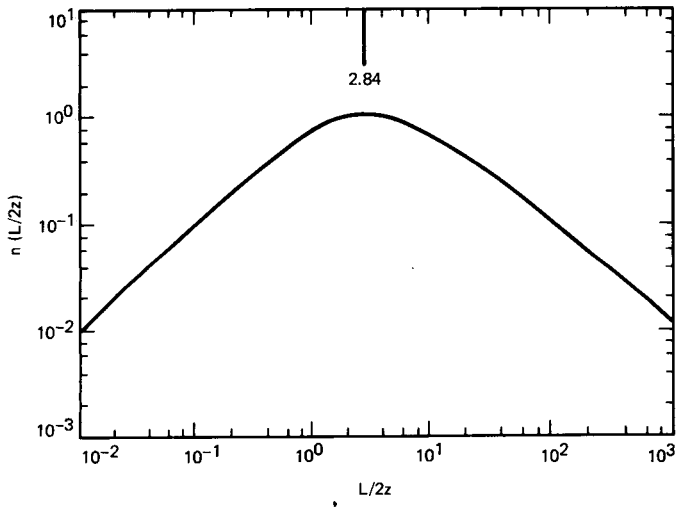


Fig. 2. Efficiency reduction factor, *h*, graphed as a function of *L/b*. Double refraction has been set to zero [4].

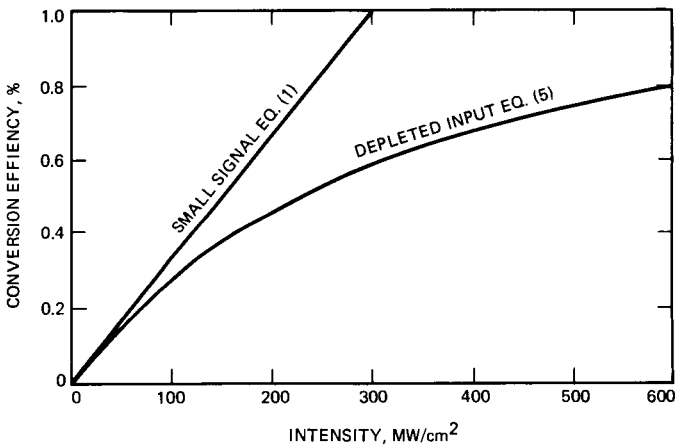


Fig. 3. Frequency doubling conversion efficiency has been graphed versus incident power density

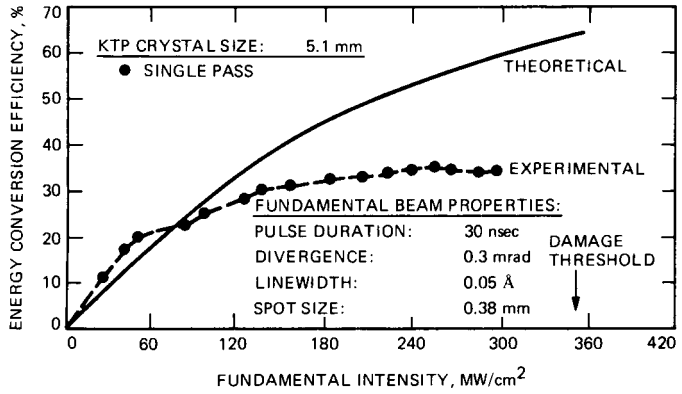


Fig. 4. Theoretical and experimental frequency doubling conversion efficiencies are plotted for single pass through a 5.1 mm KTP crystal. The saturation of the experimental results is due to reconversion and higher order processes.

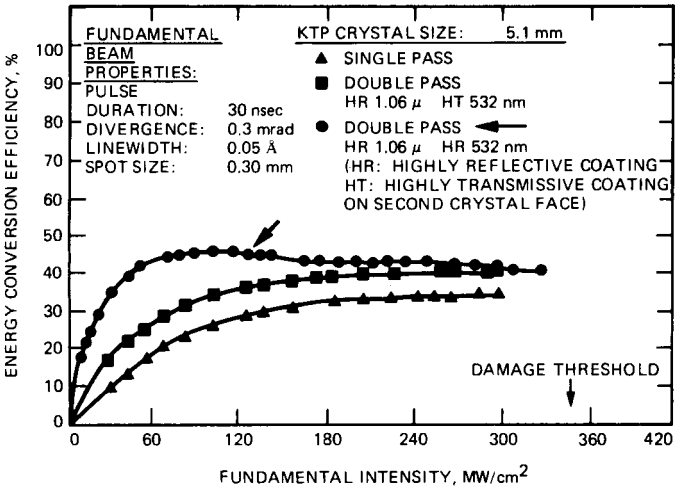


Fig. 5. Experimental conversion efficiency versus incident power density for a Gaussian 1.06 μm fundamental beam. The crystal was 5.1 mm in length [5].

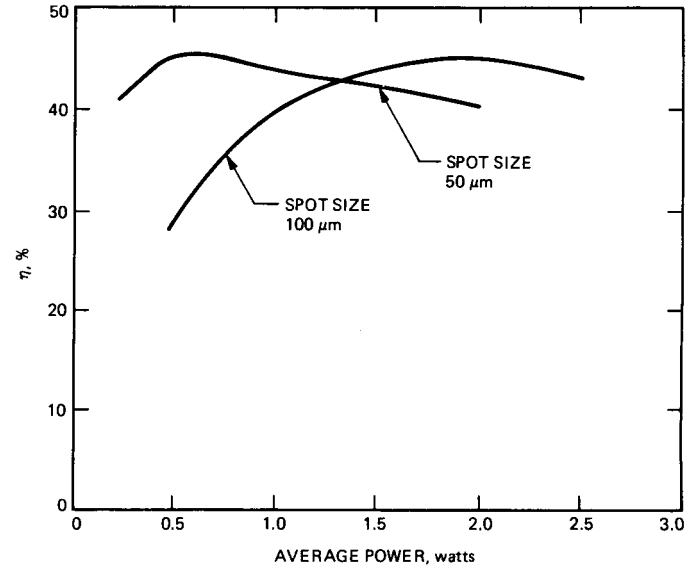


Fig. 6. Conversion efficiency versus average laser power for a PPM optical communications link. Assumed data rate is ~100 kbits with a pulse width of 20 ns and a dead time of 44.4 μs. Efficiencies for a 50 μm and 100 μm average spot size are graphed.

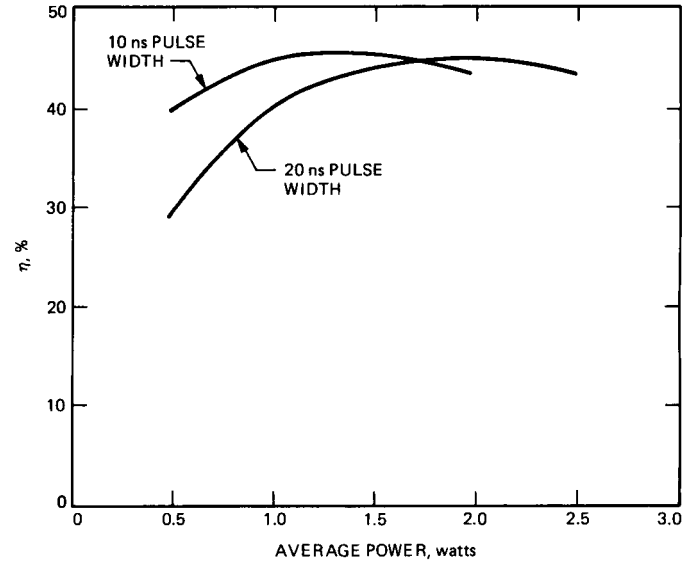


Fig. 7. Conversion efficiency versus average laser power for a PPM optical communications link. Assumed data rate is ~100 kbits with 44.4 μs dead time. Average spot size in the crystal was 100 μm. Efficiencies for pulse widths of 10 ns and 20 ns are graphed in the figure.

Appendix A
Conversion Efficiency

Crystals capable of higher order processes have nonlinear terms included in their polarization.

$$P = bE(1 + a_1E + a_2E^2 + a_3E^3 + \ldots)$$

where a_i and b are constants and E is the electric field. Specifically, let us examine the effect of the first nonlinearity, $P = dE^2$, which is responsible for second harmonic generation.

Starting with a one dimensional electric field traveling in the z direction, let us consider traveling waves of three frequencies, ω_1, ω_2, and ω_3.

$$E_1(\omega_1, z, t) = \frac{1}{2}\left[E_{1i}(z)\, e^{i(\omega_1 t - k_1 z)} + c.c.\right] \tag{A1}$$

$$E_k(\omega_2, z, t) = \frac{1}{2}\left[E_{2k}(z)\, e^{i(\omega_2 t - k_2 z)} + c.c.\right] \tag{A2}$$

$$E_j(\omega_3, z, t) = \frac{1}{2}\left[E_{3j}(z)\, e^{i(\omega_3 t - k_3 z)} + c.c.\right] \tag{A3}$$

Then, using the equation

$$\Delta^2 E = \mu_o \sigma \frac{\partial E}{\partial t} + \mu_o \epsilon \frac{\partial E^2}{\partial t^2} + \mu_o \frac{\partial^2 P_{NL}}{\partial t} \tag{A4}$$

where $(P_{NL})_i = d_{\text{eff}} E_j E_k$ is the nonlinear polarization and σ is the conductivity, it can be shown that

$$\frac{dE_{1i}}{dz} = \frac{-\sigma_1}{2}\sqrt{\frac{\mu_o}{\epsilon_1}}\, E_{1i} - \frac{i\omega_1}{2}\sqrt{\frac{\mu_o}{\epsilon_1}}\, d_{\text{eff}} E_{3j} E_{2k}^* + e^{-i(k_3 - k_2 - k_1)z} \tag{A5}$$

$$\frac{dE_{2k}^*}{dz} = \frac{-\sigma_2}{2}\sqrt{\frac{\mu_o}{\epsilon_2}}\, E_{2k}^* + \frac{i\omega_2}{2}\sqrt{\frac{\mu_o}{\epsilon_2}}\, d_{\text{eff}} E_{1i} E_{3j}^*\, e^{-i(k_1 - k_3 + k_2)z} \tag{A6}$$

$$\frac{dE_{3j}}{dz} = \frac{-\sigma_3}{2}\sqrt{\frac{\mu_o}{\epsilon_3}}\, E_{3j} - \frac{i\omega_3}{2}\sqrt{\frac{\mu_o}{\epsilon_3}}\, d_{\text{eff}} E_{1i} E_{2k}\, e^{-i(k_1 + k_2 - k_3)z} \tag{A7}$$

For second harmonic generation, $\omega_1 = \omega_2$ and $\omega_3 = \omega_1 + \omega_2 = 2\omega_1$.

If we assume that the quantity of power lost by the input beam at ω_1 due to second harmonic conversion is negligible, then

$$\frac{dE_{1i}}{dz} \cong 0$$

As a result of the above two conditions only Eq. (A7) need be examined. Furthermore, the medium chosen will usually be transparent to radiation at ω_3, hence $\sigma_3 = 0$.

Therefore, Eq. (A7) reduces to the following:

$$\frac{dE_{3j}}{dz} = -i\omega\sqrt{\frac{\mu_o}{\epsilon}}\, d_{\text{eff}} E_{1i} E_{1k}\, e^{i\Delta k z} \tag{A8}$$

where

$$\Delta k = k_3 - k_1 - k_1$$

If we take $E_{3j}(0) = 0$ for no second harmonic input and let the frequency doubler's length be L, then the solution to Eq. (A8) is

$$E_{3j}(L) = -i\omega\sqrt{\frac{\mu_o}{\epsilon}}\, d_{\text{eff}} E_{1i} E_{1k}\, \frac{e^{i\Delta k L} - 1}{i\Delta k} \tag{A9}$$

where multiplication by E_{3j}^* leads to

$$E_{3j}(L) E_{3j}^*(L) = \frac{\mu_o}{\epsilon}\omega^2 d_{\text{eff}}^2 \times E_{1i}^2 E_{1k}^2 L^2 \frac{\sin^2\left(\frac{\Delta k L}{2}\right)}{\left(\frac{\Delta k L}{2}\right)^2} \tag{A10}$$

This equation leads to an expression in terms of second harmonic output power when

$$\frac{P(2\omega)}{\text{Area}} = \frac{1}{2}\sqrt{\frac{\epsilon}{\mu_o}}\, E_{3j} E_{3j}^{*} \tag{A11}$$

is substituted into it. Thus:

$$\frac{P(2\omega)}{\text{Area}} = \frac{1}{2}\sqrt{\frac{\mu_o}{\epsilon}}\, \omega^2 (d_{\text{eff}})^2 E_{1i}^2 E_{1k}^2 L^2 \frac{\sin^2\left(\frac{\Delta kL}{2}\right)}{\left(\frac{\Delta kL}{2}\right)^2} \tag{A12}$$

Utilizing a similar expression for $P(\omega)/\text{Area}$, an expression for conversion efficiency, $P(2\omega)/P(\omega)$ can be found to be:

$$\eta = \frac{P(2\omega)}{P(\omega)} = 2\left(\frac{\mu_o}{\epsilon_o}\right)^{3/2} \frac{\omega^2 d_{\text{eff}}^2 L^2}{n^3} \frac{P(\omega)}{\text{Area}} \frac{\sin^2\left(\frac{\Delta kL}{2}\right)}{\left(\frac{\Delta kL}{2}\right)^2} \tag{A13}$$

(note that $\epsilon = \epsilon_3$ and $\epsilon_1 \cong \epsilon_3 = \epsilon_o n^2$). See main text for definition of variables.

Appendix B
Crystal Angle Phase Matching

If the crystal is birefringent, having axially dependent indices of refraction, then $k(\omega) = \omega(\mu\epsilon_0)^{1/2} n(\omega)$. The k condition of $k_2 = 2k_1$, as discussed in the main text, translates to an $n(2\omega) = n(\omega)$ condition. Since in dispersive materials n is proportional to ω, it is not possible to match $n(2\omega)$ and $n(\omega)$ if the wave is propagating down a single crystal axis, because $n(2\omega)$ will always be larger. Therefore, in order to satisfy the equality of $n(\omega) = n(2\omega)$, a combination of the crystal indices of refraction is necessary. For uniaxial crystals, two axis indices are the same: $n_x = n_y = n_o$ and $n_z = n_e$ where z is the crystal optic axis, n_o is the ordinary index of refraction and n_e is the extraordinary index of refraction. Figure B1 is a graphical representation summarizing the indices of refraction for two orthogonal polarizations at varying angular propagation directions to the optic axis for differing wavelengths. In the graphs we are considering the case of the negative uniaxial crystal, $n_o > n_e$. The wave-vector, $\mathbf{D_o}$, propagates along s with an electric field polarized normal to the page. Thus, $\mathbf{D_o}$ remains constant, as seen by the circular representation of n_o, for varying propagation angles, θ. On the other hand, D_e has an electric field polarized perpendicular to $\mathbf{D_o}$ and the direction of propagation; thus, its magnitude varies for different propagation angles, θ. Figure B1(a) illustrates these index relationships at frequency ω while Fig. B1(b) illustrates them at 2ω. Note the increased size at 2ω due to the higher frequency. In order to obtain $n(\omega) = n(2\omega)$, the 2 curves are overlapped. The intersection of the two curves yields the desired answer. The solution results from using n_o at ω and $n_e(\theta)$ at 2ω as seen in Fig. B1(c). This angular solution is analytically described as follows:

$$\frac{1}{(n_e(2\omega,\theta))^2} = \frac{\cos^2\theta}{(n_o(\omega))^2} + \frac{\sin^2\theta}{(n_e(2\omega))^2} \qquad \text{(B1)}$$

where θ is the angle between the ray's axis of propagation and the optic axis. This same type of analysis is used with biaxial crystals where $n_x \neq n_y \neq n_z$. Since all axes have different indices of refraction, the analysis is much more complicated.

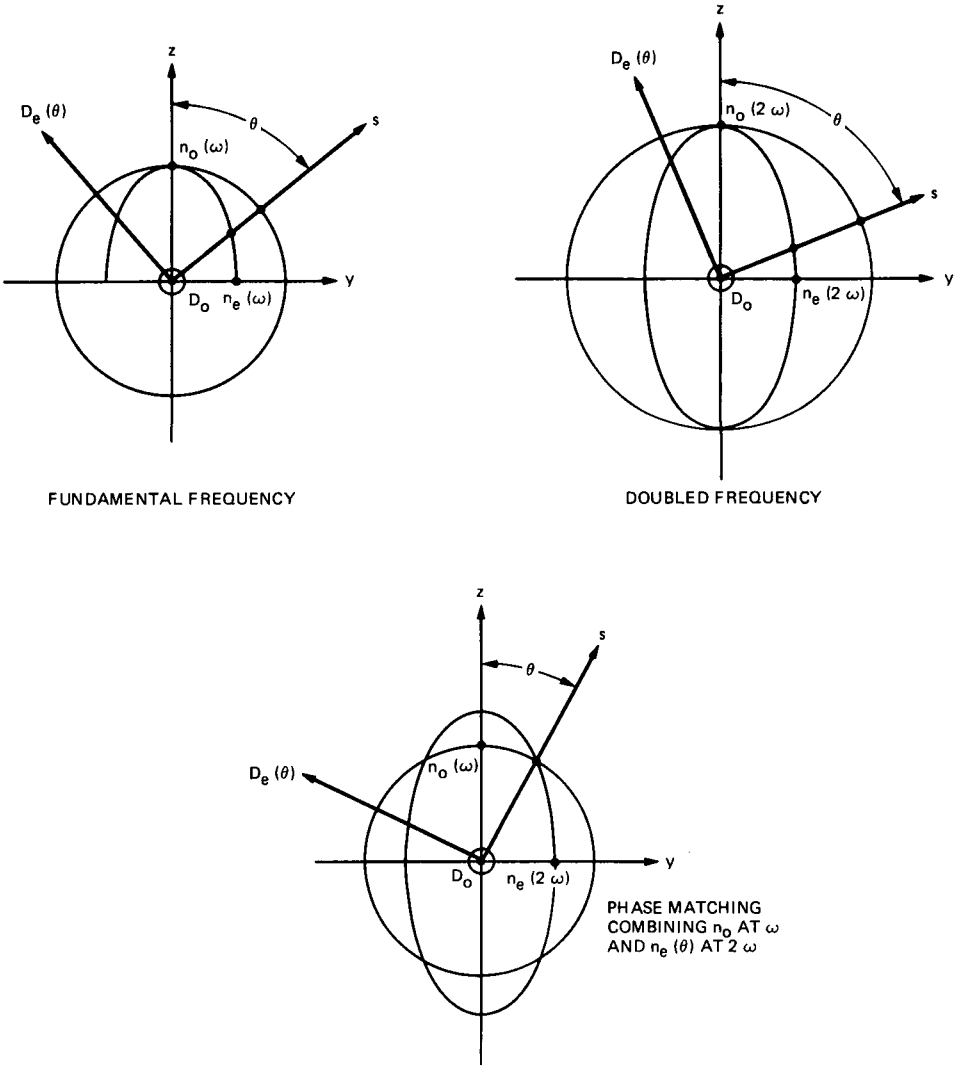


Fig. B1. Phase matching curves for ordinary and extraordinary indices of refraction

Appendix C
Depleted Input Conversion Efficiency

Let us utilize Eqs. (A5), (A6), and (A7) from Appendix A and follow a similar approach as found in Yariv [2].

$$\frac{dE_{1i}}{dz} = \frac{-\sigma_1(\mu_o)^{1/2}}{2(\epsilon_1)^{1/2}} E_{1i} - \frac{i\omega_1(\mu_o)^{1/2}}{2(\epsilon_1)^{1/2}} d_{\text{eff}} E_{3j} E^*_{2k} \epsilon^{-i\Delta kz} \tag{C1}$$

$$\frac{dE^*_{2k}}{dz} = \frac{-\sigma_2(\mu_o)^{1/2}}{2(\epsilon_2)^{1/2}} E^*_{2k} + \frac{i\omega_2(\mu_o)^{1/2}}{2(\epsilon_2)^{1/2}} d_{\text{eff}} E_{1i} E^*_{3j} e^{i\Delta kz} \tag{C2}$$

$$\frac{dE_{3j}}{dz} = \frac{-\sigma_3(\mu_o)^{1/2}}{2(\epsilon_3)^{1/2}} E_{3j} - \frac{i\omega_3(\mu_o)^{1/2}}{2(\epsilon_3)^{1/2}} d_{\text{eff}} E_{1i} E^*_{2k} e^{i\Delta kz} \tag{C3}$$

We can set $\sigma_1 = \sigma_2 = \sigma_3 = 0$ because the medium chosen should be transparent to radiation of the input frequency and the second harmonic. Furthermore, for frequency doubling, $E_1 = E_2$. Since the phases should be matched, $\Delta k = 0$. Also, if $E_1(0)$ is real, then $E_1(z)$ is real. Let us drop the $_{ijk}$ notation and define $E_3 = -iE'_3$.

Therefore, Eqs. (C1) and (C3) become:

$$\frac{dE_1}{dz} = \frac{-\omega_1 (\mu_o)^{1/2}}{2 (\epsilon_1)^{1/2}} d_{\text{eff}} E'_3 E^*_1 \tag{C4}$$

and

$$\frac{dE'_3}{dz} = \frac{\omega_3 (\mu_o)^{1/2}}{2 (\epsilon_3)^{1/2}} d_{\text{eff}} E_1^2 \tag{C5}$$

Therefore,

$$\frac{d}{dz} E_1^2 + E'^2_3 \frac{\omega_1 (\epsilon_1)^{1/2}}{\omega_3 (\epsilon_3)^{1/2}} = 0 \tag{C6}$$

If there is no input at ω_3, then integrating the above equation and substituting $\epsilon_o n^2 = \epsilon_1 \approx \epsilon_3$ yields:

$$E_1^2 + E'^2_3 \frac{(\omega_1)}{(\omega_3)} = E_1^2(0) \tag{C7}$$

Substituting E_1^2 into Eq. (C5) leads to:

$$\frac{dE'_3}{dz} = \frac{\omega_3 (\mu_o)^{1/2}}{2(\epsilon_3)^{1/2}} d_{\text{eff}} \left(E_1^2(0) - \frac{\omega_1}{\omega_3} E'^2_3 \right) \tag{C8}$$

Integration of Eq. (C8) leads to:

$$E'_3 = E_1(0) \left(\frac{\omega_3}{\omega_1} \right)^{1/2} \tanh \left(\frac{d_{\text{eff}} (\mu_o \omega_1 \omega_3)^{1/2} E_1(0) z}{2 (\epsilon_o n_1 n_3)^{1/2}} \right) \tag{C9}$$

Given that

$$\frac{P_1}{A} = \frac{n_1 (\epsilon_o)^{1/2}}{2(\mu_o)^{1/2}} |E_1|^2$$

where A = beam area, and using $\omega_3 = 2\omega_1$, Eq. (C9) becomes:

$$\eta = \frac{P(2\omega)}{P(\omega)} = \frac{|E'_3|^2}{|E_1(0)|^2} = 2 \tanh^2 \left(\frac{d_{\text{eff}} (\mu_o)^{3/4} \omega_1 P_1^{1/2} z}{A^{1/2} \epsilon_o^{3/4} n^{3/2}} \right)$$

Making the small angle approximation leads to the small signal formula of Appendix A for conversion efficiency.

The Atmosphere of Mars and Optical Communications

J. Annis
Communications Systems Research Section

The effects of the Martian atmosphere on an optical communication link are analyzed using Mariner 9, Viking Orbiter, and Viking Lander data. Clouds are found to have minimal effect because of their scarcity and thinness. Dust (from dust storms) has the dominant impact on opacity. However, periods of reduced visibility are infrequent and more closely resemble the effects of thin clouds on the Earth. A simple argument is presented which suggests that the Martian atmosphere has fewer turbulence-related effects (i.e., Mars has better resolution, lower image wander, and less scintillation) than the best of the Earth's ground-based locations.

I. Introduction

Optical communication links have been proposed for the Mars Rover. These communication systems have the high data rates and low mass, volume, and power consumption that the Rover requires. However, they face environmental problems that are different from those for radio systems, and these problems have never been evaluated for Mars and its atmosphere. This article will discuss the atmospheric effects that influence an optical communication system and will survey those effects in relation to the Martian atmosphere.

II. Optical Communications

Optical communication links are designed to operate near visible wavelengths (although they can also be in the infrared). The doubled Nd:YAG laser currently under consideration lases at 0.53 μm. This wavelength allows an exploration of the effects of the Martian atmosphere on an optical communication link using currently available data, as all the past planetary images were taken near this wavelength. In particular, the Viking Orbiters' images were at 0.44 μm to 0.59 μm and the Viking Landers' at 0.67 μm and 0.59 μm.

The two major effects of an atmosphere on an optical communication link are attenuation and scintillation [9]. Attenuation is caused by absorption and scattering due to both air molecules and aerosols. Optical depth, τ, is a measure of attenuation over the entire path length, here taken to be the distance from the ground to space. Optical depth increases as the line of sight moves down toward the horizon, increasing the path length. The power received, P_r, is the power transmitted, P_t, multiplied by the attenuation: $P_r = P_t e^{-\tau}$. An optical depth of 1 attenuates a signal by 63 percent. For comparison, a clear night on Earth has an optical depth of $\approx$0.2. The effect of attenuation is to reduce the visibility of the link.

Scintillation is caused by turbulence-induced spatial and temporal refractive index variations. Refractive index variations also cause the wavefront distortions that determine

resolution and image wander. The effect of scintillation is to increase the bit error rate. Resolution and image wander affect beam-pointing accuracy.

There may be some concern that aerosol dust affects resolution on Mars. There are really a number of problems involved in that concern. The laser beam from the surface may suffer an increased beam divergence due to the aerosols, or the laser beacon/illuminated Earth's image may be so far underresolved that pointing the laser is difficult. Both of these issues have been examined experimentally. Doubled Nd:YAG lasers fired through terrestrial fog with optical depths of 10 or greater suffered no increased beam divergence [8]. The same experiments found an increase in the apparent angle of the source (say, the Earth) by a factor of 4 at $\tau \approx 2$. This does not seem to pose a problem, as the typical optical depth is more like 0.5 and the base resolution of the atmosphere is high (see Section V). Comparison of fog with dust is justified because both are primarily Mie scatterers. This aspect of the dust, as well as a new problem that this brings, will be examined next.

Martian aerosol dust behaves as a Mie scatterer, except that it has a tendency to scatter more at high angles than does a Mie scatterer [6]. Mie scattering is characterized by large amounts of forward scattering. A receiver with a wide enough field of view would see three concentric circles: the direct (but attenuated by the optical depth) beam, the multiple forward-scattered photons, and the diffused photons (see Fig. 1, based on a figure in [5]). The light from both of the last two circles comes from the attenuation of the direct beam. It is possible for the multiple forward-scattered light, which has a size on the order of 2 degrees, to overwhelm the direct beam. While this does not concern a link looking at a laser fired through an aerosol (it just receives more photons than it should), the ability of a link to pick out the beacon/Earth might be in jeopardy. If the forward component dominated, the link would see a blur with a size of about 2 degrees. However, just by energy conservation, the scattered component cannot even equal the direct beam intensity until $\tau \approx 0.7$ ($e^{-0.7} \approx 0.5$), and experiments show that this does not occur until an optical depth of 15 [5]. At optical depths of 1 this does not seem to be a substantial concern.

III. The Atmosphere

On Mars, scattering by atmospheric molecules is negligible ($\tau \approx 0.002$), mainly because the atmosphere is extremely thin. Molecular absorption is also unimportant outside of a few deep absorption bands. This is also the case on Earth, where absorption features have been extensively tabulated between 0.4 μm and 10.0 μm. The locations and relative strengths of the Martian atmospheric absorptions are the same as those of the Earth's atmosphere to a first approximation, because most of the same gases appear in about the same proportions in each. On Mars, CO_2 will have relatively stronger lines and H_2O relatively weaker, but the only major addition, CO, has no absorption lines below 2.3 μm. These lines are easily avoided. Overall, there will be a pronounced decrease in the strength of the atmospheric absorption lines on Mars as a result of the thin atmosphere.

IV. The Aerosols

The major attenuators on Mars are the aerosol scatterers: clouds, fog, haze, and dust. The global and local properties of the attenuators have been studied using Mariner 9, Viking Orbiter, and Viking Lander data. Orbiter data are down-looking images that allow cloud type, cloud distribution, and optical depth to be determined. The Viking Lander data are up-looking images of the Sun and Phobos that allow precise optical depth measurements.

A. Combined Effects

The optical depth of the Martian atmosphere and its variation over a season were examined by Thorpe [10] using Viking Orbiter images taken during northern summer/southern winter. The measurements were made in three regions: the Viking Lander 1 region (northern equatorial latitudes); the Smooth Plains (northern equatorial region and mid-latitudes); and the Old Terrain (southern mid-latitudes). The Viking Lander 1 region results are consistent with those from the lander itself. The other two regions have optical depth histories that are similar to that of the Lander 1 site: $\tau \approx 0.2$–0.3, with short excursions to 0.6. These probably represent the thicker hazes that Kahn's statistics (see Subsection C) say should be in at least one place in these regions 20 to 30 percent of the time.

The Viking Landers returned optical depth data at two locations on the surface of Mars for almost a Martian year. The history at 22 degrees N, Viking 1 in the northern equatorial latitudes, is shown in Fig. 2. The history at 48 degrees N, Viking 2 in the northern mid-latitudes, is shown in Fig. 3. (Both figures are from [6], and both express the time coordinate in terms of sol number, i.e., the number of Martian solar days from touchdown.) The global dust storms are the two excursions to high optical depths. Other than the dust storm, the largest effect seen was at Viking Lander 2, where for less than a day the optical depth grew larger than 1, reaching 2.8 (see Fig. 4, derived from [11]). This was identified as a cold front associated with the north polar hood passing the lander [11]. It was not seen at the other lander farther south. For comparison, a terrestrial cumulus cloud has an optical depth of $\approx$ 6–200. Viking 1 found a background haze of $\tau \approx 0.3$, while Viking 2 found a similar haze of $\tau \approx 0.5$. Pollack *et al.* [6], [7] identify this as 5 μm dust particles from local dust storms

suspended in the atmosphere. Both landers also saw a diurnal variation in τ of about 0.2 (see Fig. 5, derived from [6]) during the Martian summer. As this additional optical depth appears before sunrise and disappears around noon, it has been identified with (H_2O) fog.

B. Dust

Local dust storms were seen to occur almost entirely in two locations: at the edge of the retreating southern polar cap and in the low-latitude regions of the southern hemisphere [7]. The coverage of the southern hemisphere was more complete than that of the northern hemisphere, however. The low-latitude storms occur mostly when Mars is at perihelion. Local dust storms last on the order of a day.

Global dust storms can last for 70 days. They typically happen once a Martian year, in the southern hemisphere's late spring to summer. The Martian year that the Viking Landers spent on Mars was particularly bad; there were two global dust storms. The more southerly lander, Viking 1, saw higher optical depths measuring $\tau = 2.7$, and $\tau = 3.7$ for the storms. Pollack *et al.* [7] estimate an upper limit on the dust storms of $\tau = 3.2$ and $\tau = 9$, respectively. These storms are apparently caused by solar heating of the atmosphere acting on local dust storms in a feedback mechanism. They feature high winds and blowing dust, so it is doubtful that any communication link would be used during a storm.

C. Clouds

The Viking Orbiters and Mariner 9 provided us with 58,000 images of Mars. Of these, 2.4 percent had some form of cloud or fog, and 28 percent had visible haze. These images have been cataloged and the cloud occurrence statistics tabulated by Kahn [4]. Table 1 shows a simplified version of this table. Although this grossly oversimplifies the weather patterns of the planet, the statistics can help sketch out the large patterns. The statistics are broken up by cloud type, season, and latitude range. Each entry should be read as "the probability of finding at least one cloud of this type in this latitude range during this season." Not surprisingly, there is a distinct seasonal dependence as well as a latitude dependence. Not shown are a longitude dependence and a time-of-day dependence [3], [4]. Some Martian clouds form at dawn and burn off rapidly, and others form only in the midday.

In general terms the occurrence patterns of Martian clouds may be described as follows, paraphrased from Kahn [4]. The northern hemisphere is, in general, more cloudy than the southern hemisphere in corresponding seasons. Clouds generally form more easily in mid-latitudes than in the equatorial regions, and more easily still in the polar regions. Clouds are relatively abundant during northern spring and summer at mid-latitudes; in the southern hemisphere mid-latitudes the situation is complicated by atmospheric dust. Thick H_2O clouds are not found at high latitudes during mid-to-late autumn or winter in either hemisphere, nor are they found in southern mid-latitudes during early winter. Overall, the optical thickness of Martian clouds is $\approx$0.05–3.0, a figure closer to terrestrial cirrus clouds ($\tau \approx 0.3$–3.5) than to stratus clouds ($\tau \approx 6$–80) or cumulus clouds ($\tau \approx 5$–200) [9].

Widespread, optically thick clouds occur in three regions [1]. The polar regions have a seasonally dependent bank of H_2O and CO_2 clouds as well as hazes around their perimeters, known as polar hoods. It seems the hazes have optical thicknesses on the order of 1, while clouds are thicker [2]. The second region is the Tharsis Bulge. Here, on the northwestern flanks of the four large volcanoes, optically thick clouds ($\tau \geqslant 1$) are found perhaps 25 percent of the time. These clouds are caused by wind moving over the volcanoes. The third area of widespread clouds is the plateau region south of the western end of the Marineris valley, where a field of cellular clouds is a daily occurrence during the summer.

Isolated clouds are often found in other locations [1], [3], [4]. The most common type is the lee cloud, which can be optically thick. These fairly small clouds form downwind of craters, in the same way as the Tharsis Volcano clouds. When they occur, these clouds are numerous enough to be used to map out wind patterns on a regional scale. Wave clouds are much like lee clouds in thickness and in occurrence rates but are not associated with a ground obstacle. Optically thin cirrus clouds are relatively common. These clouds, probably CO_2, are widespread when they occur but have optical thicknesses of only $\approx$0.05. Hazes are also noted by Kahn [4]. It is not clear if they are dust or ice, but they can have optical depths $\geqslant 1$.

D. Fog

Optically thick fog was seen in some crater bottoms and in the Marineris valley, particularly in the Labyrinthus Noctus [1]. Terrestrial fogs are about 150 meters thick and have optical depths of $\approx$3. Although the hourly coverage was not good, the fog seems to burn off in the afternoon. The fog seen by the Viking Landers was thin, about $\tau = 0.2$, and apparently would not have been seen by the orbiters. Fog occurs mostly in the spring and summer, and mostly in the southern hemisphere.

V. Turbulence

The size of the turbulence-induced effects is hard to determine from Viking data. The images give practically no information, as the expected effects are below the resolution limits of the imagers (the 0.12-degree diodes on the landers were

stepped in 0.04-degree intervals; on Earth bad resolution is 0.03 degree). However, the theories of index-of-refraction effects developed for the Earth's atmosphere can be applied to the case of the Martian atmosphere. An order-of-magnitude argument using these theories is presented here.

The distortion capabilities of turbulence are directly related to the size of the mean square fluctuations of the index of refraction, C_N. In turn, C_N is directly related to the size of the mean square fluctuations of temperature, C_T, with the relation expressed as:

$$C_N = \text{const} \frac{P}{T^2} C_T$$

where T is the temperature and P is the pressure. The ratio of the fluctuations on Mars to those on Earth is then:

$$\frac{C_{N,\,\text{Mars}}}{C_{N,\,\text{Earth}}} = \frac{P_{\text{Mars}}}{P_{\text{Earth}}} \left(\frac{T_{\text{Earth}}}{T_{\text{Mars}}}\right)^2 \frac{C_{T,\,\text{Mars}}}{C_{T,\,\text{Earth}}}$$

The first factor on the right is on the order one hundredth, and the second factor on the right is on the order of 1. It seems very unlikely that the temperature fluctuations on Mars are one hundred times larger than those on Earth (bear in mind that these are short time scale variations). More likely is that the index of variation fluctuations in the Martian atmosphere is one hundred times smaller than that of the Earth's atmosphere. This suggests that the average location on Mars has a view of an optical communication link that is one hundred times more stable than the average location on Earth.

VI. Conclusions

Mars's weather may be simpler than the Earth's weather, but it is still a very complex system. By making sweeping generalizations, the overall results can be shown in Table 2, which compares Mars and the Earth. In general, optically thick clouds and fog are fairly rare on Mars. Aerosol dust is always present and is the main source of atmospheric opacity. The great dust storms are the cause of the highest opacity, $\tau \approx 10$. Attenuation losses due to clouds, fog, and dust seem modest. However, such visibility still corresponds to thin clouds on the Earth. The turbulence-induced effects, such as scintillation and wavefront distortion, should be less of a problem on Mars than on Earth. Aerosol dust may have some impact on resolution, but it should not be great. More than likely, the surface of Mars has routinely better visibility than Earth's best locations both in terms of resolution and in terms of lack of scintillation and image movement. It certainly has clear weather much more often.

References

[1] G. Briggs, K. Klaasen, T. Thorpe, J. Wellman, and W. Baum, "Martian Dynamical Phenomena During June-November 1976: Viking Orbiter Results," *J. Geophys. Res.*, vol. 82, pp. 4121-4149, 1977.

[2] P. R. Christensen and R. W. Zurek, "Martian North Polar Hazes and Surface Ice: Results from the Viking Survey/Completion Mission," *J. Geophys. Res.*, vol. 89, pp. 4587-4596, 1984.

[3] R. G. French, P. J. Gierasch, B. D. Popp, and R. J. Yerdon, "Global Patterns in Cloud Forms on Mars," *Icarus*, vol. 45, pp. 468-493, 1981.

[4] R. Kahn, "The Spatial and Seasonal Distribution of Martian Clouds and Some Meteorological Implications," *J. Geophys. Res.*, vol. 89, pp. 6671-6688, 1984.

[5] G. C. Mooradian, M. Geller, L. B. Stolts, D. H. Stephens, and R. A. Krautwald, "Blue-Green Propagation Through Fog," *Applied Optics*, vol. 18, pp. 429-441, 1979.

[6] J. B. Pollack, D. Colburn, R. Kahn, J. Hunter, W. Van Camp, C. E. Carlston, and M. R. Wolf, "Properties of Aerosols in the Martian Atmosphere, as Inferred from Viking Lander Imaging Data," *J. Geophys. Res.*, vol. 82, pp. 4479-4496, 1977.

[7] J. B. Pollack, D. S. Colburn, F. M. Flasar, R. Kahn, C. E. Carlston, and D. Pidek, "Properties and Effects of Dust Particles Suspended in the Martian Atmosphere," *J. Geophys. Res.*, vol. 84, pp. 2929–2945, 1979.

[8] W. S. Ross, W. P. Jaeger, J. Nakai, T. T. Nguyen, and J. H. Shapiro, "Atmospheric Optical Propagation: An Integrated Approach," *Applied Optics*, vol. 21, pp. 775–785, 1982.

[9] The Technical Cooperation Program, "The Application of Optical Space Communication to the Military Communications Requirements: A User's Guide," presented at the Laser Communication Workshop, Technical Panel STP-6, Space Communications, Salisbury, Australia, October 29 to November 2, 1984.

[10] T. Thorpe, "Viking Orbiter Observations of Atmospheric Opacity During July–November 1976," *J. Geophys. Res.*, vol. 82, pp. 4151–4159, 1977.

[11] J. E. Tillman, R. M. Henry, and S. L. Hess, "Frontal Systems During Passage of the Martian North Polar Hood over the Viking Lander 2 Site Prior to the First 1977 Dust Storm," *J. Geophys. Res.*, vol. 84, pp. 2947–2955, 1979.

Table 1. Martian cloud occurrence statistics

Season	Latitude range	Occurrence probabilities (%)			
		Thin clouds ($\tau < 1$)	Thick clouds ($\tau \geq 1$)	Fog	Haze ($\tau \geq 1$)
Northern spring	Polar	9	24	0	22
	Mid-latitudes	6	10	0	10
	Equatorial	6	6	0	8
Southern fall	Equatorial	5	16	5	7
	Mid-latitudes	3	19	3	8
	Polar	26	59	8	7
Northern summer	Polar	31	24	7	27
	Mid-latitudes	11	10	0	13
	Equatorial	10	16	3	24
Southern winter	Equatorial	8	12	3	10
	Mid-latitudes	19	23	17	15
	Polar	4	4	4	0
Northern fall	Polar	12	1	0	60
	Mid-latitudes	11	6	0	38
	Equatorial	1	1	0	25
Southern spring	Equatorial	6	9	5	33
	Mid-latitudes	10	27	18	54
	Polar	28	7	33	67
Northern winter	Polar	7	0	0	0
	Mid-latitudes	10	17	0	32
	Equatorial	2	2	2	13
Southern summer	Equatorial	17	7	6	19
	Mid-latitudes	19	26	28	21
	Polar	33	18	21	30

Table 2. Comparison of Earth and Mars generalized weather patterns

Atmospheric condition	Earth		Mars	
	Typical optical depth	Distribution	Typical optical depth	Distribution
Atmospheric absorption and Rayleigh scattering	0.2	Everywhere	0.002	Everywhere
Aerosol dust	–	–	0.5	Everywhere
Dust storms	–	–	10.0	Southern hemisphere or global
Fog	≈3	Many places	≈1.0	Morning; crater bottoms and valleys
			0.2	Morning; everywhere?
Clouds H_2O	≈5	50% cloud cover	≈1.0	Winter polar; isolated; clouds behind high places
Clouds CO_2	–	–	≈0.001	Many places
			≈1.0	Winter polar regions

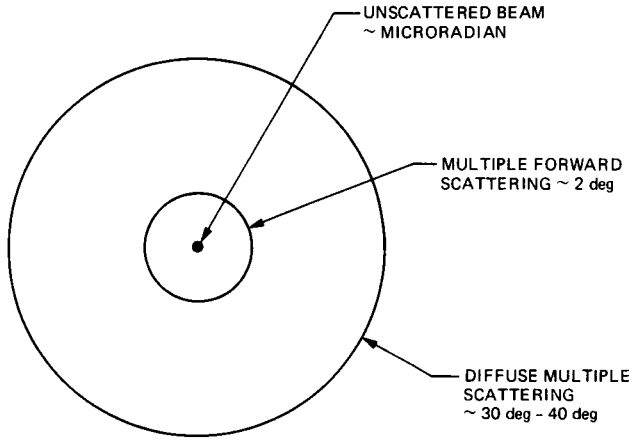


Fig. 1. Schematic representation of laser angular brightness distribution due to Mie scattering in optically thick regions

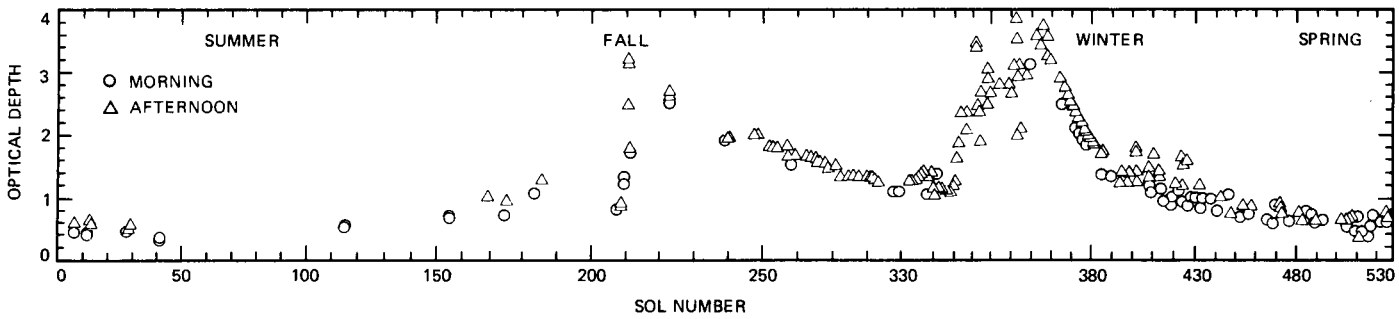


Fig. 2. Optical depth at the Viking Lander 1 site as a function of time

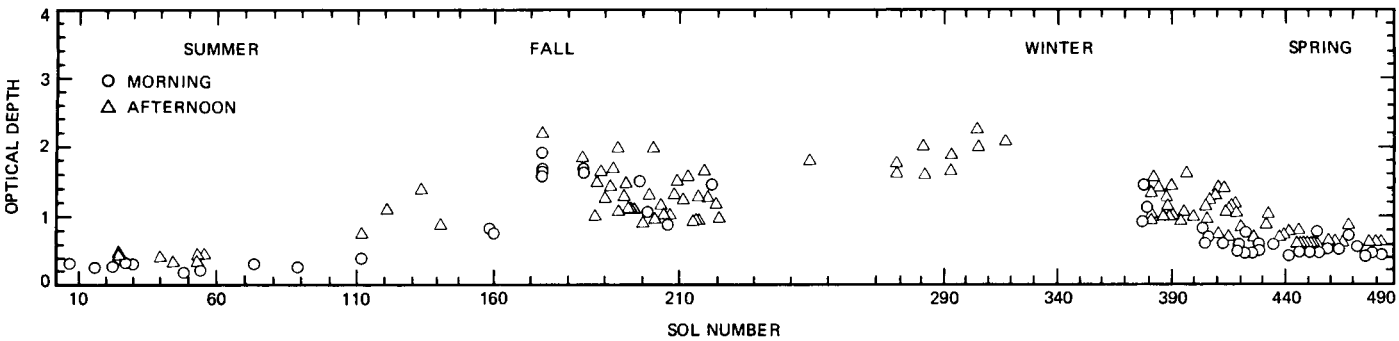


Fig. 3. Optical depth at the Viking Lander 2 site as a function of time (sol 0 here is equivalent to sol 44 for the Viking Lander 1)

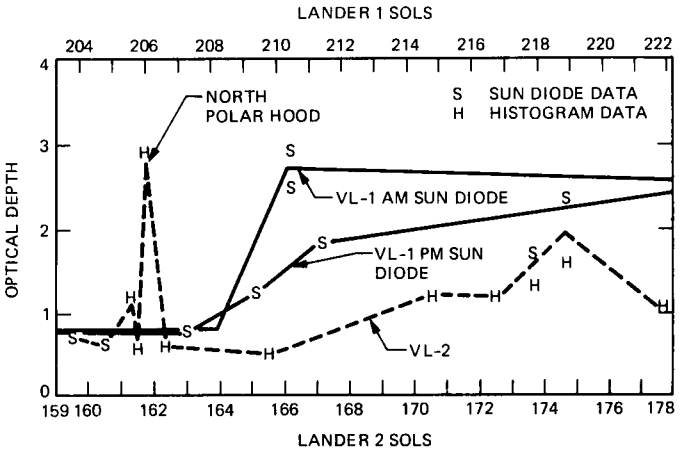


Fig. 4. Optical depth data from both landers (histogram data are derived from analysis of images, while the sun diode data are derived from imaging the sun)

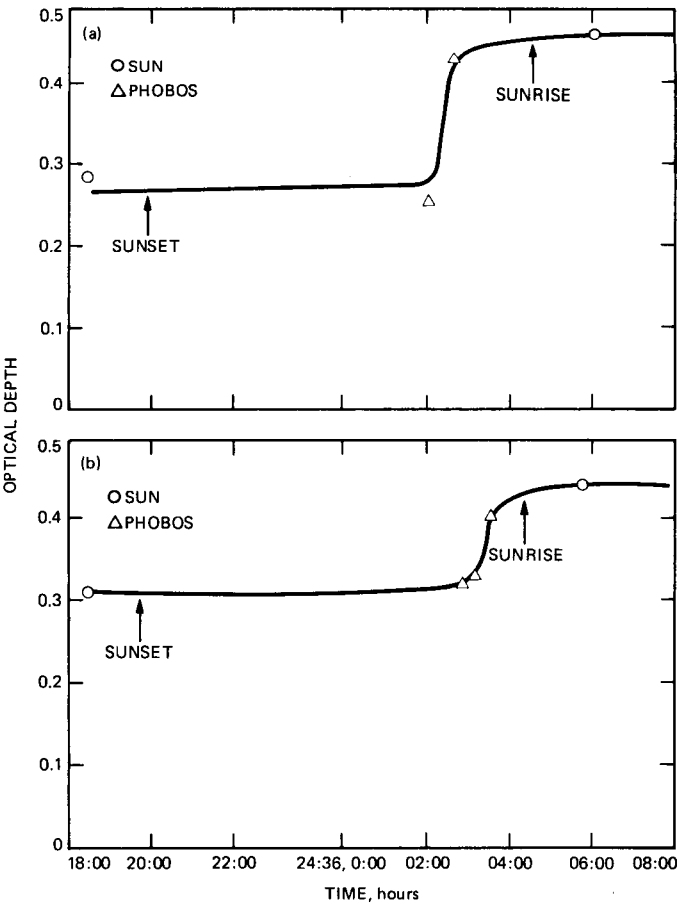


Fig. 5. Variation of optical depth (a) from the late afternoon of sol 24 to the early morning of sol 25 at the Viking Lander 2 site, as inferred from photograph of Phobos (triangles) and the sun (circles); (b) similar data from the late afternoon of sol 28 to the early morning of sol 29

A Near-Earth Optical Communications Terminal With a Corevolving Planetary Sun Shield

E. L. Kerr
Communications Systems Research Section

The umbra of a planet may serve as a sun shield for a space-based optical communications terminal or for a space-based astronomical observatory. An orbit that keeps the terminal or observatory within the umbra is desirable.

There is a corevolution point behind every planet. A small body stabilized at the planet corevolution point will revolve about the Sun at the same angular velocity as the planet, always keeping the planet between itself and the Sun. This corevolution point is within the umbra of Mars but beyond the end of the umbra for Mercury, Venus, and Earth.

The Mars corevolution point is an ideal location for an astronomical observatory. There Mars obstructs less than 0.00024 percent of the sky at any time, and it shades the observatory completely from the Sun. At the Earth corevolution point, between 51 percent and 84 percent of the solar disk area is blocked, as is up to 92 percent of the sunlight. This provides a reduction from 3 dB to 11 dB in sunlight that could interfere with optical communications if scattered directly into the detectors. The variation is caused by revolution of the Earth about the Earth–Moon barycenter.

Corevolution is also possible closer to the planet than the corevolution point, but only if continuous outward thrust is provided or if a tether is attached to another body located beyond the corevolution point. Corevolution close enough to the Earth to make the Earth disk appear 75 percent larger in diameter than the solar disk (in order to block sunlight at any position of the Earth and Moon) is practically unobtainable, since it would require a tether 900,000 km long, or 5.6 newtons (1.3 pounds) of continuous outward thrust for a 10,000-kg terminal or observatory.

I. Introduction: Natural Sun Shields

Planetary bodies are usually used as sun shields for astronomical observations. On Earth we look at the stars at night. There are four inherent difficulties: the atmosphere of the Earth interferes with visibility; the gravity of the Earth distorts the optics; the bulk of the Earth and the thickness of the atmosphere cut the view to about three-eighths of the heavens at any one time at night; and scattered sunlight during the day obscures the stars. The great advantage of Earth basing is ease of access to the telescope.

An Earth-based optical communications terminal for deep-space data links will operate by day as well as by night. It will filter out most sunlight interference with a narrowband filter. An additional reduction of sunlight interference can be achieved by tuning the communications lasers to a solar Fraunhofer line [1]. Long sunshades, some with internal, venetian-blind slats across the aperture, will permit viewing within a small angle θ of the center of the Sun. The Sun will therefore obstruct an additional $(1 - \cos\theta)/2$ of the sky. For $\theta = 5$ degrees, the solar obstruction is 0.19 percent.

Inaccessibility of five-eighths of the sky at any given time would require placement of a network of three terminals around the Earth for continuous coverage. More terminals would be required in order to provide reasonable assurance of visibility at any given time. Otherwise, the telescope in the terminal suffers the same difficulties as an Earth-based astronomical observatory telescope.

A lunar-based observatory or optical communications terminal would not have to contend with the atmosphere. Observing and optical communication would be possible by day as well as by night, since there would be no sky scattering. Gravitational distortion would be much less severe. Half the heavens would still be obstructed at any one time by the bulk of the Moon. The disk of the Sun has a half-angle of one-quarter degree. The solar obstruction of the sky would be only a little more than 0.0005 percent. However, sunlight scattering within the telescope and heating of the telescope would present more severe difficulties for daytime observing than they would for an Earth-based telescope. Thermal cycling would be gradual, since both the lunar day and the lunar night are over two weeks long. Continuous sky coverage by a lunar optical communications network could be achieved by two terminals on opposite sides of the Moon.

A space-based telescope is troubled neither by atmospherics nor by gravity. In low planetary orbit, somewhat less than half the sky is obscured at any one time by the bulk of the host planet. The obscured portion changes rapidly as the telescope revolves around the planet. Thermal cycling can be as rapid as 45 minutes each of heating and cooling. Higher orbits reduce the amount and rate of movement of the sky obstruction caused by the planet. They lengthen the thermal cycle and increase the ratio of heating to cooling in each cycle. The solar sky obstruction and internal light scattering remain difficulties.

There is an ideal location in space, at the corevolution point within the umbra of a planet, where most of these difficulties are removed. The sky obstruction of the host planet and the Sun may be consolidated into one small obstruction which is the larger of the two. This location will now be defined, described, and analyzed.

II. Definitions and Descriptions

A. Planetary Umbral Points

As a planet p revolves around the Sun, its umbra is a cone-shaped region coming to a point a fixed distance directly behind the planet from the Sun, as shown in Fig. 1. At this point, called the umbral point, the angular subtense of the Sun and that of the planet are exactly equal. The distance U_p from the planet center to the umbral point is set by the radii of the planet r_p and of the Sun $r_\odot$, and by the distance R of the planet from the Sun:

$$U_p = R\frac{r_p}{r_\odot - r_p}$$

At the umbral point, the Sun and planet disks are the same size. Exactly 100 percent of the area of the solar disk is blocked by the planet, and no more of the sky is obstructed by the planet than the amount that would be obstructed by the Sun. At any other distance D from the planet center along the line pointing directly away from the Sun, the percentage of the solar disk area obscured by the planet is

$$\left[\left(\frac{r_p}{r_\odot}\right)\left(\frac{D+R}{D}\right)\right]^2 \times 100\%$$

B. The Corevolution Point

The amount of centripetal acceleration required to keep a body moving in a circular path around a central object is v^2/R, where v is the tangential velocity of the body. Solar gravitation, which supplies the available centripetal acceleration, becomes weaker inversely as the square of R, so the outer planets must move more slowly than the inner planets in order to stay in circular orbits.

On a line directly behind a planet from the Sun, the available centripetal acceleration is augmented by the gravitation

of the planet. There is a point on this line, called the corevolution point, at which the total gravitational pull is just enough to keep a small body moving around the Sun at the same angular velocity as the planet. Closer to the planet the pull is too strong, so the small body will accelerate inward; farther from the planet the pull is weaker, so the body will accelerate outward. In either case, the body no longer revolves about the Sun at the same angular rate as the planet, and the orbit is no longer circular.

If the small body stays at the corevolution radius from the Sun but gets ahead of or behind the planet, then the gravitational forces are no longer colinear, and the body will decelerate or accelerate appropriately. Likewise, if the body moves above or below the planetary orbital plane without changing its distance from the Sun, the noncolinear forces will bring it back. In short, the corevolution position is a saddle point of equilibrium for the body. It is unstable in the radial direction but stable in the two angular directions.

The distance C_p of the corevolution point from the planet center depends on the masses of the planet and of the Sun and on the distance from the planet to the Sun. The distance is calculated in the theory section of this article. The stability is analyzed in [2].

C. Relationship Between the Umbral and Corevolution Points

Both the umbral and corevolution points are located on the line directly behind the planet from the Sun, but their relative distances depend on the radius and mass of the planet, respectively. A very dense planet will have its corevolution point beyond the umbra, while a light, fluffy planet will have an umbra reaching beyond the corevolution point. This is illustrated in Fig. 1 for Earth, with a density of 5588 kg/m^3, and for Mars, with a density of 3968 kg/m^3. As a consequence, the Sun as seen from the Earth corevolution point is a thin ring (sometimes complete and sometimes broken as a result of the revolution of the Earth and the Moon about their barycenter), while Mars completely blocks the view of the Sun from its corevolution point.

III. Uses of Corevolution Points

A corevolution point is perpetually shaded by the host planet from most or all sunlight. Furthermore, the two sky obstructions caused by the planet and the Sun are consolidated into the larger of the two, with the rate of movement reduced to the planetary revolution rate. There are two important applications.

A. An Ideal Astronomical Observatory Location

An astronomical observatory would ideally be located at a corevolution point that happened to be just a little closer to the planet than the umbral point. Of the inner planets, only Mars provides an ideal location. The corevolution point is 1.08 million kilometers beyond the center of Mars. At that distance, the disk of Mars blocks out the Sun and only 5 percent more area of the sky. The excess diameter of Mars should be sufficient to compensate for any fluctuations due to the ellipticity of the Mars orbit and for refraction through the thin Martian atmosphere. The observatory would require a radioisotope thermoelectric generator or another electrical power source, since no sunlight is available at the location for solar power.

B. A Good Location for a Near-Earth Optical Communications Terminal

The corevolution point for the Earth lies 1.5 million kilometers beyond the Earth. The umbral point is closer to the Earth, so the terrestrial disk area is only 84 percent of the solar disk area. This location is useful for optical communications. There is sufficient sunlight to provide electrical power.

The Earth is the only planet with a natural satellite of significant mass relative to its own mass. The lunar mass is 1.2 percent of the terrestrial mass. The Earth and the Moon revolve about their own barycenter, and thus the Earth, as seen from a point directly behind it from the Sun, oscillates in the plane of the lunar orbit with an amplitude of 4671 km. As seen from the corevolution point, the Earth blocks 51 percent of the solar disk at the extremes of its motion and 84 percent as it passes through the center. The variation in heating and cooling is not sinusoidal, so the average value differs somewhat from the median value of 67.5 percent reduction in heating with swings of ±16.5 percent. The thermal period is half the lunar revolution period—a little over two weeks.

Solar interference with communications is also reduced geometrically by 3 dB to 8.2 dB when light is scattered directly into the telescope detectors from the sun shades or optics. A further reduction is provided by the limb-darkening effect, in which the apparent surface of the Sun appears to have reduced brightness at the edges relative to the center. The effect is more pronounced as the wavelength becomes smaller. At a wavelength of 534 nm, the solar intensity drops to 79 percent at 0.75 of the solar radius from the center and to 55 percent at 0.95 of the solar radius from the center [3]. The terrestrial disk reaches out to 0.92 of the solar radius from the center (when aligned with the Sun). If the integrated intensity reduction over the visible solar annulus is taken to be 50 percent, limb darkening provides an additional 3-dB reduction of

solar interference whenever the Earth disk is centered directly over the Sun.

The position of the communications terminal will be slightly perturbed by the orbiting of the Earth and the Moon about their barycenter. The amount of thrust and the quantity of energy required for station keeping have yet to be analyzed in detail, but they are expected to be small. The mass of the Moon is only 1.2 percent of the Earth's mass, and the radius of the Moon's orbit is only one-quarter of the corevolution distance. These circumstances reduce the Moon's influence considerably.

A deep-space probe will be able to locate the near-Earth optical communications terminal more easily if the terminal is at the corevolution point rather than in an Earth-centered orbit. Once the probe locates the Sun and the Earth, it has two points that establish a line. The corevolution point is always the same distance along the line from Earth away from the Sun. The probe must compute only the foreshortening of that distance according to the epoch and its own position relative to the ecliptic. In contrast, if the terminal is in an Earth-centered orbit, the probe still needs accurate information on six more orbital elements to locate the terminal after it has located the Earth and the Sun.

IV. Corevolution Theory

A small body placed behind a planet on the line from the Sun through the planet would experience centripetal acceleration from both the planet and the Sun. Close behind the planet, the centripetal acceleration is much greater than the acceleration of the planet toward the Sun. As the distance from the planet increases, centripetal acceleration decreases. At a certain point on the line, herein called the corevolution point, the centripetal acceleration is sufficient to keep the body moving in a circular path around the Sun at exactly the same angular rate as the planet. (Technically speaking, there are two other corevolution points. One lies between the planet and the Sun, while the other lies on the opposite side of the Sun from the planet. Neither point has any relevance to the present discussion.)

A. The Restricted Three-Body Problem

Consider three masses placed on a straight line and revolving steadily about their common center of mass. Let the bodies be arranged in order of decreasing mass, with the mass of the third body negligible compared with the masses of the other two. Let R be the distance between the first two, and let C be the distance between the last two. The center of mass will lie between the first two bodies at a distance $X = Rm_2/(m_1 + m_2)$. The square of the angular revolution rate is $G(m_1 + m_2)/R^3$, where G is the universal gravitational constant.

The amount of centripetal acceleration required to keep the third mass moving in a circular path is the square of the angular revolution rate times the distance to the center of mass, $R + C - X$. This centripetal acceleration is supplied by gravitation from the first two masses and may be reduced by an outward thrust or tether tension T directed away from the two massive bodies. Dividing Newton's law of forces and accelerations for the third body by G and m_3 yields

$$\frac{m_1}{(R+C)^2} + \frac{m_2}{C^2} - \frac{T}{Gm_3} = \frac{m_1 + m_2}{R^3}(R + C - X)$$

Let $x = C/R$ and let $M = m_1/m_2$. If T is set to zero, one may solve the following quintic equation for x:

$$f(x) = 1 + 2x + x^2 - [(1 + 3M)x^3 + (2 + 3M)x^4 + (1 + M)x^5] = 0$$

B. Calculations

The quintic equation is easily solved using Newton's method for the approximation of roots. The value of M is large for any planet and $x \ll 1$, so an initial approximation for x is $(m_2/3m_1)^{1/3}$. Only a few refinements are required to find x when the planet has the combined masses of the Earth and Moon and the Earth's distance from the Sun. Sample iterations of the solution are shown in Table 1 using Sun, Earth, and Moon data found in Table 2. The distance C is 1.5073×10^6 km. This point is in the penumbra. Geometrically, 84.20 percent of the area of the Sun's disk is blocked by the Earth. These calculations may be performed with a SuperCalc program called COREV.CAL, written by the author.

C. Stability

Steady solutions have been sought for the motion of three bodies placed colinearly and revolving about the common center of mass. This configuration is stable in the angular directions but not in the radial direction. A corevolution point is therefore a saddle point, and we may expect to find no natural body at the corevolution point of any planet. However, an active system could stabilize itself at a corevolution point, expending energy and fuel only for station-keeping purposes to compensate for small perturbations from natural satellites or distant planets. The thrusts required are expected to be small.

D. Tabulation for All of the Planets

The corevolution distance C from planet center to satellite is given in Table 2 for each planet on the basis of planet mass and planet mean radius R from the Sun. The mass of a planet is usually much larger than the total mass of its natural satellites, and the natural satellite masses may be neglected. Earth is the only exception recognized here. The relative distance x is computed first from the quintic equation using Newton's method.

The percentage block is the area ratio of the planet to solar disk as seen from the corevolution point. This area ratio (if less than 100 percent) may be used to give a geometric reduction factor in the solar flux on the satellite. The flux will actually be reduced further by solar limb darkening. This further reduction is due to absorption in the solar atmosphere on slant paths near the limbs. It depends on the wavelength and is difficult to calculate exactly. Block percentages greater than 100 percent represent the amount of sky obscured by the planet relative to the solar obscuration at the satellite location.

E. Corevolution at a Distance Closer than the Corevolution Point

If one desires to place a satellite closer to the Earth in order to be completely within the umbra, one may choose the value of x and solve for the additional constant thrust or outward tension required to make the satellite corevolve. Suppose, for example, x is chosen to be 0.0112597 so that the apparent diameter of the Earth will be 75 percent greater than the apparent diameter of the solar disk. The distance xR now corresponds to 786,556 km. The thrust or tension required for a 10,000-kg satellite is 5.60 newtons (1.26 pounds). This amount seems small, but there is no practical way to provide it (see Appendix A).

V. Conclusion

Mars is the best sun shield for a space telescope. A telescope placed at the Mars corevolution point would experience no sunlight interference or thermal cycling and could view all but 0.00024 percent of the sky at any time.

The Earth corevolution point is a good location for the near-Earth end of an optical communications data link to deep-space probes. It is an easy point to locate. Thermal cycling of the telescope is reduced by 51 percent to 84 percent, and interference from scattered sunlight is reduced from 3 dB to 11 dB. All but from 0.00057 percent down to 0.00053 percent of the sky is visible there at any time.

References

[1] E. L. Kerr, "Fraunhofer Filters to Reduce Solar Background for Optical Communications," *TDA Progress Report 42-87*, vol. July–September 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 48–55, November 15, 1986.

[2] K. R. Symon, *Mechanics* (3rd ed.), Reading, Massachusetts: Addison-Wesley, pp. 286–291 and 490–497, 1971.

[3] G. Abetti, *The Sun*, New York: Macmillan, pp. 276–283, 1957.

[4] W. A. Baracat and C. L. Butner, *Tethers in Space Handbook*, Washington, D. C.: National Aeronautics and Space Administration, August 1986.

Table 1. Sample iterations of the solution of the quintic equation when finding the Earth corevolution point

n	x_n	$f(x_n)$	$x_n - x_{n-1}$
0	0.0100442	0.0101104	
1	0.0100778	−0.000034	0.0000336
2	0.0100777	-3.9×10^{-10}	-1.14×10^{-7}
3	0.0100777	-2.2×10^{-16}	-1.3×10^{-12}
4	0.0100777	-2.2×10^{-16}	0

Table 2. Corevolution distance and solar blockage for all of the planets (solar and lunar data are also included)

Planet	m, kg	r, km	R, km	x	C, km	Block, %
Mercury	3.181×10^{23}	2433	5.795×10^{7}	0.0037670	2.183×10^{5}	86.77
Venus	4.883×10^{24}	6053	1.081×10^{8}	0.0093795	1.014×10^{6}	87.61
Earth	6.053×10^{24}	6371	1.496×10^{8}	0.0100777	1.507×10^{6}	84.20
Mars	6.418×10^{23}	3380	2.278×10^{8}	0.0047616	1.085×10^{6}	105.02
Jupiter	1.901×10^{27}	69,758	7.781×10^{8}	0.0697847	5.430×10^{7}	236.10
Saturn	5.684×10^{26}	58,219	1.427×10^{9}	0.0463373	6.612×10^{7}	356.82
Uranus	8.682×10^{25}	23,470	2.870×10^{9}	0.0246016	7.061×10^{7}	197.27
Neptune	1.027×10^{26}	22,716	4.500×10^{9}	0.0260302	1.171×10^{8}	165.53
Pluto	1.256×10^{22}	1100	5.909×10^{9}	0.0012817	7.574×10^{6}	152.47
Sun	1.991×10^{30}	695,950				
Moon	7.354×10^{22}	1738	3.844×10^{5}	(from the Earth)		

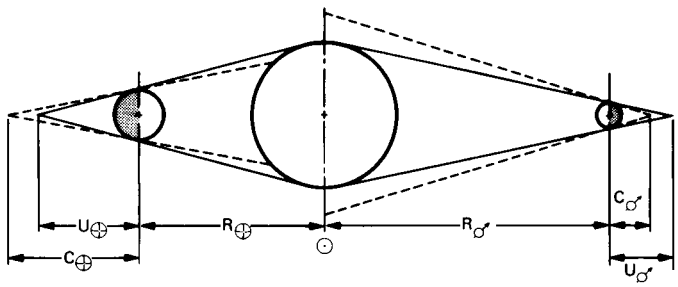


Fig. 1. Umbral and corevolution points of Earth (left) and Mars (right) in relation to the Sun (the relative diameters of Earth and Mars are drawn to scale, and their relative distances from the Sun are shown to a different scale; all other distances are exaggerated or reduced for clarity). Solid lines bound umbras; broken lines project planet limbs.

Appendix A
Corevolution Within the Earth Umbra

If it were desirable to locate an optical communications terminal entirely within the Earth umbra and still have it corevolve with the Earth, one might do so in two ways: by providing a constant thrust away from the Earth or by tethering the observatory to another body located beyond the corevolution point. The second body would be in the penumbra, where sufficient light would be available to generate power and send it to the optical communications terminal via a pair of wires that would serve as the tether.

The second option is not practical because of the length and mass of the tether. Suppose, for example, that the optical-communications-terminal mass is 10,000 kg and the solar-power generator is ten times more massive. (Additional mass at the generator would be required to balance the mass of the tether, as its center of mass does not coincide with the corevolution point.) The observatory may be located 790,000 km from the Earth's center, where the Earth subtense is 75 percent more than the solar subtense. The corevolution point is at 1.51 million kilometers, and the generator is located at 1.68 million kilometers. The tension in the pair of wires would be 5.60 newtons (1.26 pounds). A steady direct current in the wires would maintain their separation by the magnetic field, and the voltage would be limited only by the insulation and vacuum gap at the connection points. Even so, the length of 898,000 km would make the mass of two aluminum wires 0.5 mm in diameter equal to 813,000 kg. The taper required for constant stress is very slight and therefore unnecessary. However, the resistance of the wires would be very large even in the cold of space unless a relatively high temperature extrudable superconducting material were found for them.

Alternatively, the power could be supplied by other means and the tether could be a plastic filament. If the filament diameter were 0.12 mm and the density were 1450 kg/m^2, the filament mass could be reduced to 14,000 kg, still very large.

A number of space applications for tethers are identified in [4]. The longest tether proposed there is 526,000 km long, so the present suggestion may be a record. The length of this tether by itself would require the listing of its "potential for technology demonstration" as "far-term."

The first option is also impractical. Supplying even a fraction of an ounce of thrust for one year would require a mass of fuel about equal to the mass of the optical communications terminal.

Optimized Tracking of RF Carriers With Phase Noise, Including Pioneer 10 Results

V. A. Vilnrotter, W. J. Hurd, and D. H. Brown
Communications Systems Research Section

The ability to track very weak signals from distant spacecraft is limited by the phase instabilities of the received carrier signal and of the local oscillator employed by the receiver. These instabilities ultimately limit the minimum loop bandwidth that can be used in a phase-coherent receiver, and hence limit the ratio of received carrier power to noise spectral density which can be tracked phase-coherently. This article presents a method for near-real time estimation of the received carrier phase and additive noise spectrum, and optimization of the phase locked loop bandwidth. The method was used with the breadboard DSN Advanced Receiver to optimize tracking of very weak signals from the Pioneer 10 spacecraft, which is now more distant than the edge of the solar system. Tracking with bandwidths of 0.1 Hz to 1.0 Hz reduces tracking signal threshold and increases carrier loop SNR by 5 dB to 15 dB compared to the 3 Hz bandwidth of the receivers now used operationally in the DSN. This will enable the DSN to track Pioneer 10 until its power source fails near the end of the century.

I. Introduction

The ability to track very weak signals from distant spacecraft is limited by the phase instabilities of the received carrier signal and of the local oscillator employed by the receiver. These instabilities ultimately limit the minimum loop bandwidth that can be employed in a phase-coherent receiver, and hence limit the ratio of received carrier power to noise spectral density which can be tracked phase-coherently. To achieve the best performance in a telemetry system, the spectral density of the phase noise should be known, and the parameters of the phase locked loop should be optimized.

At the present time, the Deep Space Network (DSN) uses two distinct operational telemetry receivers. These receivers, designated Block III and Block IV, respectively, employ second order analog carrier tracking loops, with a minimum loop bandwidth of 3 Hz. A 1 Hz bandwidth is possible with the Block IV receiver when tracking 2.3 GHz (S-band) signals, but this bandwidth is seldom used due to receiver oscillator noise, and because of the relatively large phase error caused by the inability of narrow bandwidth second order loops to track dynamics introduced by the rotation of the earth.

One of the important advantages of digital phase locked loops over their analog counterparts is in the flexibility and accuracy of setting and modifying loop parameters. In the DSN Advanced Receiver, which is currently in breadboard form, the loop filter characteristics can be altered in real time. This capability motivates the search for an adaptive algorithm to optimize loop parameters, particularly loop bandwidth, in

real time. Optimization is especially important when tracking weak signals generated by noisy oscillators, such as the signal currently being received from Pioneer 10. When the loop is not stressed by dynamics, or equivalently when the loop order is high enough to track dynamics with negligible phase error, then the optimum loop bandwidth may be defined as that bandwidth which minimizes the steady state rms phase error within the loop. At its optimum value, the loop bandwidth is typically wide enough to track out phase jitter in the received carrier, yet narrow enough to effectively limit the amount of thermal noise allowed into the loop. By reducing the carrier phase jitter, radio loss is reduced and data recovery improved.

In order to optimize loop bandwidth, it is first necessary to model the phase noise properly. The significant components of phase noise are due to the spacecraft transmitter, propagation effects, and VCO noise at the receiver. Spacecraft transmitter noise depends on the mode of transmission, that is, one-way, two-way, or three-way. With one-way transmission, a free running oscillator on the spacecraft determines the transmitted phase. With two- and three-way transmission, the spacecraft phase locks to an uplink signal, multiplies its frequency and phase by a fixed ratio, and uses the resultant as the downlink carrier. If the downlink receiver and uplink transmitter are at the same station, then the mode is termed two-way, while the term "three-way transmission" is reserved for the case when the downlink receiver is at a different station. In the absence of propagation effects, one-way reception is usually dominated by spacecraft oscillator noise, whereas uplink transmitter noise normally dominates two- and three-way reception. A significant propagation effect is that due to solar scintillation, caused by charged particles in the vicinity of the Sun. This effect dominates when the spacecraft is more distant than the Sun, and the Sun-earth-probe (SEP) angle is small. Clearly, all of the above effects depend on a variety of factors; hence any real time characterization of the phase process can be of great value during reception. In this article we present a method for estimating the total phase spectral density in a phase locked receiver, and specifically in the DSN Advanced Receiver. The technique involves measuring the spectral density at the phase detector output, and extracting the individual spectral components by processing the estimated received spectrum. Spectral estimates are obtained by the application of fast Fourier transforms (FFTs) to the phase detector output. Given knowledge of the closed loop transfer function, the phase spectrum, receiver noise spectrum, and received carrier power can all be estimated. The optimum loop bandwidth is determined from these estimated components. Analytic expressions are derived for the case when the phase noise spectral density varies as $f^{-\alpha}$, a model often quoted in the literature. With α equal to three, the model can be used to describe oscillator phase spectra at low frequencies, while for solar scintillation spectra a value of $\alpha = 8/3$ is appropriate.

Finally, we present measurements of phase spectra, and determine optimum loop bandwidths for Pioneer 10 telemetry. Data for both one-way and three-way transmissions are included. It is shown that the Advanced Receiver can reduce the minimum tracking threshold for Pioneer 10 by as much as 5 dB to 15 dB compared to the tracking threshold of current operational receivers. This is accomplished by using 0.1 Hz to 1.0 Hz loop bandwidths, compared to the 3 Hz bandwidths currently used operationally.

II. Mathematical Models

In the following development, we assume that the received signal is of the form

$$s(t) = \sqrt{2}\, A \cos(\omega_0 t + \theta(t) + \psi)$$

where A is the signal amplitude, ω_0 is the carrier radian frequency, $\theta(t)$ is a random phase process, and ψ is a random initial phase. Multiplication by $\sqrt{2}$ has the advantage that the total carrier power becomes $P_c = A^2$, and the slope of the phase detector S-curve near the origin becomes A. This signal representation also allows direct comparison between the baseband phase spectral densities within the phase locked loop and their RF counterparts, where oscillator phase spectra are generally measured. Since in the current application the signal amplitude remains essentially constant for long periods of time (and hence can be estimated with great accuracy), we assume that A is known.

A. Derivation of Observation and Phase Error Spectra

A linear model of the carrier loop is shown in Fig. 1. This model is accurate when the loop is locked, and operating with small rms phase errors. Under these conditions, the loop is continuously generating estimates of the received phase process $\theta(t)$, denoted $\hat{\theta}(t)$, and subtracting the estimate from the received phase. The resulting phase error process $\phi(t) = \theta(t) - \hat{\theta}(t)$ is multiplied by the slope of the phase-detector S-curve at zero frequency (in our case that value is "A"), and filtered by the loop filter in the presence of additive receiver noise $n(t)$. The output of the loop filter is an error process $e(t)$ which is used to control the frequency of the voltage-controlled oscillator (VCO), or the equivalent numerically controlled oscillator (NCO) typically employed in digital loops. Instabilities within the receiver VCO are modeled as an additive phase process $\psi_r(t)$ that is added to the VCO output. At this point it is useful to introduce the Heaviside differential operator $p = d/dt$, which allows relating the output "z" of a linear filter to its input "x" in the time domain as $z(t) = F(p)\, x(t)$, where $F(p)$ is a ratio of polynomials in p. The

transfer function of the filter in the frequency domain becomes $F(j\omega) = F(p)|_{p=j\omega}$. (Subsequently we shall use both radian frequency ω and frequency $f = \omega/2\pi$, as appropriate.) Suppressing time dependence and making use of the Heaviside differential operator, the loop equations can be expressed as

$$\phi = \theta - \hat{\theta} \tag{1a}$$

$$y = A\phi + n \tag{1b}$$

$$x = KG\left\{\frac{F(p)}{p}\right\} y \tag{1c}$$

$$\hat{\theta} = x + \psi_r \tag{1d}$$

where K is the filter gain, G the VCO gain, and y the noise-corrupted observable. In general, the received phase consists of phase processes due to modulation, doppler, and transmitter VCO instabilities. Any initial phase offsets are assumed to have been tracked out by the receiver. For the type of signals typically observed by the Advanced Receiver, the use of subcarriers can be assumed, which implies that modulation sidebands are far removed from the carrier and hence may be ignored. Therefore, the received phase can be modeled as $\theta = d + \psi_t$ where d is due to possible doppler rates and ψ_t describes the effects of additive transmitter VCO instabilities. Defining the transceiver phase process $\Delta\psi = \psi_t - \psi_r$ and letting $B(p) = KGF(p)/p$, Eqs. (1a), (1c), and (1d) can be used to obtain

$$\hat{\theta} = B(p)\,y + \psi_r = AB(p)\,\phi + B(p)\,n + \psi_r \tag{2a}$$

$$\phi = d + \Delta\psi - AB(p)\,\phi - B(p)\,n \tag{2b}$$

or

$$\phi = \left\{\frac{1}{1+AB(p)}\right\}(d+\Delta\psi) - \left\{\frac{AB(p)}{1+AB(p)}\right\}\frac{n}{A} \tag{3a}$$

$$y = A\left\{\frac{1}{1+AB(p)}\right\}(d+\Delta\psi) + \left\{\frac{1}{1+AB(p)}\right\} n \tag{3b}$$

Since the closed-loop transfer function of the loop is

$$H(j\omega) = \frac{AB(j\omega)}{(1+AB(j\omega))} \tag{4}$$

the power spectral densities of the observation and phase error processes can be represented as

$$S_\phi(f) = |1 - H(j2\pi f)|^2\, S_\zeta(f) + |H(j2\pi f)|^2\, \frac{S_n(f)}{A^2} \tag{5a}$$

$$\frac{S_y(f)}{A^2} = |1 - H(j2\pi f)|^2\, S_\zeta(f) + |1 - H(j2\pi f)|^2\, \frac{S_n(f)}{A^2} \tag{5b}$$

where $\zeta = d + \Delta\psi$. If the doppler and transceiver phase processes are independent, then $S_\zeta(f) = S_d(f) + S_{\Delta\psi}(f)$. Note that in Eq. (5b) we divided both sides of the equation by A^2 in order to emphasize the similarities between the observation and phase error spectra, when both are expressed in the same units. Aside from a scaling factor, the only difference between the two equations is that the additive noise component in the phase error process is filtered by $H(j2\pi f)$, whereas in the observation process it is effectively filtered by $(1 - H(j2\pi f))$.

Now let us consider the received phase process, specifically in the case where d is small. The received phase spectrum then depends on the transmitted phase spectrum and on the properties of the channel. The free space channel is generally considered to be an additive white Gaussian noise channel which does not introduce phase distortion. For now we will neglect the effects of the sun and the solar corona on the received phase spectrum, as these will be discussed in a later section. Therefore, in this case the phase stability of the received carrier is dominated by the phase stability of the transmitted carrier. Two transmission modes commonly used in the DSN will be examined in this analysis.

In the one-way mode, the spacecraft oscillator is the source of the transmitted carrier. The term one-way stems from the notion of the frequency reference traveling only one way, from the spacecraft to the ground station. In the one-way mode, the phase noise spectrum is normally dominated by a term varying as inverse f^3, for frequencies very close to the carrier.

The other transmission mode discussed here is known as three-way transmission. In this mode, a carrier reference is transmitted to the spacecraft from a ground station. The spacecraft receiver phase locks to this signal, multiplies the frequency by a known ratio, and uses the resultant as the downlink frequency. Thus the received phase depends on the phase characteristics of the uplink oscillator and of the spacecraft phase locked loop. In our experiments with the Pioneer 10 spacecraft, transmission in the two-way mode is not feasible since the round trip light time is roughly 11 hours 30 minutes, so that by the time the transmitted reference is returned

by the spacecraft, the transmitting antenna is on the other side of the world.

For three-way telemetry, the received process spectrum cannot be described by a simple inverse f^3 model, in general. Now the effects of the transmitter tracking loop on the uplink spectrum must be properly taken into account. A linear model of the tracking loop on the spacecraft appears in Fig. 2. This model is similar to the one describing the receiver tracking loop, except that now the loop's output is its estimate of the uplink phase. The subscript "t" denotes "transmitter," distinguishing the components of this loop from those of the receiver. The transmitter loop tracks the uplink phase, using it as the reference for the downlink telemetry. Thus, the phase of the downlink carrier is identical to the loop's estimate of the uplink phase. The power spectral density of the downlink transmitted phase process θ_t can be derived with the same techniques used to derive the spectral density of the receiver phase process. The result is

$$S_{\theta_t}(f) = |H_t(j2\pi f)|^2 \left[S_{\theta_u}(f) + \frac{S_{n_t}(f)}{A^2} \right]$$

$$+ |1 - H_t(j2\pi f)|^2 \; S_{\psi_t}(f) \qquad (5c)$$

where θ_u is the random phase of the uplink oscillator multiplied by the spacecraft transponder ratio. Note that the uplink phase process and the internal noise are both filtered by the closed loop transfer function, while the transmitter's phase process is effectively filtered by the complementary filter function $(1 - H(j2\pi f))$. In our case $S_{\theta_u} \cong 10^{-3}\, S_{\psi_t}$, but $|1 - H_t(j2\pi f)|^2 \lesssim 10^{-4}$ at the low frequencies of interest, so that typically S_{θ_u} is the dominant component of the downlink phase process.

The transmission channel may also affect the spectral purity of the carrier phase. A good example of a prominent channel effect is provided by the solar corona [1]. The resulting phase degradation can be appreciable, often negating the benefits of highly stable earth-based oscillators operating in three-way mode. The power spectrum of the phase degradation for this solar effect is typically inversely proportional to the 8/3 power of frequency.

B. Closed Loop Transfer Function and Its Effects

Next we examine the structure of the closed-loop transfer function in greater detail, and consider its effects on various phase process and additive noise models.

The Advanced Receiver employs second and third order loops. Approximating the rapidly sampled data loop by a continuous time loop, the transfer functions of the loop filter and of the resulting closed loop for a third order, type III loop are [2]

$$F(s) = \frac{1 + \tau_2 s}{\tau_1 s} + \frac{1}{\tau_2 \tau_3 s^2} \qquad (6a)$$

$$H(s) = \frac{r(k + \tau_2 s + (\tau_2 s)^2)}{r\,(k + \tau_2 s + (\tau_2 s)^2) + (\tau_2 s)^3} \qquad (6b)$$

where τ_1, τ_2, τ_3 are constants, $r = Ak\tau_2^2/\tau_1$, $k = \tau_2/\tau_3$ and s is the Laplace transform variable. The steady-state transfer function is obtained by letting $s = j\omega$ in Eq. (6b), yielding

$$H(j\omega) = \frac{r\left(k - \frac{\omega^2 g^2}{B_L^2}\right) + \frac{j\omega g r}{B_L}}{r\left(k - \frac{\omega^2 g^2}{B_L^2}\right) + j\left(\frac{\omega r g}{B_L} - \frac{\omega^3 g^3}{B_L^3}\right)} \qquad (7a)$$

$$1 - H(j\omega) = \frac{-j\,\frac{\omega^3 g^3}{B_L^3}}{r\left(k - \frac{\omega^2 g^2}{B_L^2}\right) + j\left(\frac{\omega r g}{B_L} - \frac{\omega^3 g^3}{B_L^3}\right)} \qquad (7b)$$

where

$$g \triangleq B_L \tau_2 = \frac{r}{4}\left(\frac{r - k + 1}{r - k}\right) \qquad (8)$$

These transfer functions can also be expressed in terms of the normalized radian frequency $\tilde{\omega} = \omega / B_L$, indicating that B_L simply scales the frequency variable. Note that the loop reduces to a second order type II loop when $k = 0$. The squared magnitudes of the transfer functions defined in Eq. (7) are of fundamental importance in determining loop performance. These functions are

$$F_1(\tilde{\omega}) \triangleq |H(j\omega)|^2_{\omega = B_L \tilde{\omega}} = \frac{r^2(k - \tilde{\omega}^2 g^2) + \tilde{\omega}^2 (gr)^2}{r^2(k - \tilde{\omega}^2 g^2)^2 + (\tilde{\omega} r g - \tilde{\omega}^3 g^3)^2} \qquad (9a)$$

$$F_2(\tilde{\omega}) \triangleq |1 - H(j\omega)|^2_{\omega = B_L \tilde{\omega}} = \frac{\tilde{\omega}^6 g^6}{r^2(k - \tilde{\omega}^2 g^2)^2 + (\tilde{\omega} r g - \tilde{\omega}^3 g^3)^2} \qquad (9b)$$

Plots of these functions are shown in Fig. 3 for $r = 4$, and $k = 0.0$ and 0.25. Note that for $k \leq 0.5$ and $r \geq 2$, $F_2(\tilde{\omega})$ is a monotone increasing function of $\tilde{\omega}$, which implies that $B_{L2} > B_{L1}$, $F_2(\omega/B_{L1}) > F_2(\omega/B_{L2})$. This property can be used to establish the existence of an optimum loop bandwidth.

C. Oscillator Phase Noise

Oscillator phase noise can be modeled as [3]

$$S_\psi(f) = \begin{cases} \dfrac{S_1}{|f|^3} + \dfrac{S_2}{f^2} + S_3 & |f| \leq F_u \\ \dfrac{S_4}{f^4} & |f| > F_u \end{cases} \tag{10}$$

where F_u is an upper frequency cutoff and the S_i, $i = 1, 4$, are constants. For our interest, we observe that if the transmitter and receiver oscillators are running independently, then their spectral densities add, yielding $S_{\Delta\psi}(f) = S_{\psi_t}(f) + S_{\psi_r}(f)$. Each component is typically of the form given in Eq. (10), although the corresponding constants may be different for the two oscillators. The low-frequency behavior of an oscillator is usually dominated by "frequency flicker noise" [4], giving rise to the well known inverse f^3 behavior of phase fluctuations. Assuming that in the region of interest frequency flicker noise dominates, and further assuming negligible doppler contribution, the power spectral density of the phase process for one-way telemetry can be modeled as

$$S_\xi(f) \cong \frac{S_1}{|f|^3} \tag{11}$$

This model may also be applied to three-way telemetry, if the spacecraft loop bandwidth is so great that the uplink phase process term dominates.

At low frequencies the additive receiver noise is generally white. Denoting its single-sided spectral level by N_0 yields the representation

$$S_n(f) \cong \frac{N_0}{2} \tag{12}$$

With these approximations, the process, additive noise, and observation spectral densities appear as in Fig. 4, for a third order loop with parameters as before.

Note that for filter functions of the form defined in Eq. (7) and inverse f^3 phase processes, the peak of the process spectrum always occurs at a frequency that is a linear function of the loop bandwidth. This can be established by differentiating $|1 - H(j\omega)|^2 S_\xi(f)$ with respect to ω and setting the result equal to zero. The maximum occurs at frequency $f = f^*$, where

$$f^* = C_0(r, k) B_L \tag{13}$$

(the derivation is shown in the Appendix). For ($r = 4, k = 0.25$), $f^* = 0.127\ B_L$; for a second order loop with parameters ($r = 4$, $k = 0.0$), $f^* = 0.147\ B_L$. This result can be used to establish whether or not an observed phase process is adequately described by the f^{-3} model.

D. Optimization of Tracking Loop Bandwidth

The performance of phase-locked loops can be assessed in terms of the total rms phase error σ_ϕ present in the loop. The total mean-squared phase error is the integral of the phase error spectral density defined in Eq. (5a):

$$\sigma_\phi^2 = 2 \int_0^\infty S_\phi(f)\, df \triangleq \sigma_\xi^2 + \sigma_n^2 \tag{14}$$

The components of the total phase error σ_ξ^2 and σ_n^2 are defined as

$$\sigma_\xi^2 = 2 \int_0^\infty |1 - H(j\,2\pi f)|^2 S_\xi(f)\, df \tag{15a}$$

and for white noise spectra of the form $S_n(f) = N_0/2$,

$$\sigma_n^2 = 2\left(\frac{N_0}{2A^2}\right) \int_0^\infty |H(j\,2\pi f)|^2\, df = \left(\frac{N_0}{2A^2}\right) 2B_L \tag{15b}$$

Note that for white noise the spectral level is constant; hence the component of the variance due to white noise increases linearly with loop bandwidth. The process component depends on the power spectrum of the phase, which is generally not constant. However, since the power spectral density is a nonnegative function of frequency, it follows that for $B_{L2} > B_{L1}$, $S_\xi(\omega)\, F_2(\omega/B_{L1}) > S_\xi(\omega)\, F_2(\omega/B_{L2})$. Since this inequality also holds for the integral, the component of the variance due to process noise is a decreasing function of loop bandwidth, for all valid power spectral density functions. It follows from the above argument that a loop bandwidth exists which minimizes the total mean squared phase error.

Loop bandwidth optimization can be achieved for arbitrary phase spectra by computing estimates of the total phase error as a function of loop bandwidth, and selecting the bandwidth

corresponding to the minimum of this function. For the case of additive white receiver noise, this requires knowledge of the parameter $N_0/2A^2$, and of the function $|1 - H(j\,2\pi f)|^2\, S_\xi(f)$, which is the integrand in Eq. (15a). If the power spectral density of the observable and the density of the signal level A are known, then the required integrand can be determined from Eq. (5b):

$$|1 - H(j\,2\pi f)|^2\, S_\xi(f) = \frac{S_y(f)}{A^2} - |1 - H(j\,2\pi f)|^2 \left(\frac{N_0}{2A^2}\right) \tag{16}$$

A major simplification in the optimization algorithm results if the phase process spectrum is assumed to be of the form $f^{-\alpha}$. This assumption is normally valid for one-way transmission, and may be accurate for three-way transmission as well, provided that the instability is dominated either by uplink oscillator or solar scintillation phase noise. Introducing the change of variables $\tilde{f} = f/B_L$, it follows that

$$\sigma_\xi^2 = 2\,S_1 B_L^{(1-\alpha)} \int_0^\infty |1 - H(j\,2\pi\tilde{f})|^2\, \tilde{f}^{-\alpha}\, d\tilde{f}$$

$$= 2\,S_1\, \gamma_0(\alpha) B_L^{(1-\alpha)} \tag{17}$$

where $\gamma_0(\alpha)$ is the value of the integral. Using numerical integration, we found that $\gamma_0(3) = 9.08$ and $\gamma_0(8/3) = 5.87$ for a third order loop with parameters $r = 4$, $k = 0.25$ and a loop bandwidth of 1 Hz.

For the $f^{-\alpha}$ model, the minimum of the total phase error variance can be found by differentiating σ_ϕ^2 with respect to B_L, setting the result equal to zero, and solving for the optimum loop bandwidth B_L^*:

$$\frac{\partial \sigma_\phi^2}{\partial B_L} = -2(\alpha - 1)\, S_1\, \gamma_0(\alpha) B_L^{-\alpha} + 2\left(\frac{N_0}{2A^2}\right) = 0 \tag{18a}$$

yields

$$B_L^* = \left\{ \frac{(\alpha - 1)\, S_1\, \gamma_0(\alpha)}{\dfrac{N_0}{2A^2}} \right\}^{1/\alpha} = \left\{ \frac{(\alpha - 1)\, \sigma_\xi^2}{\sigma_n^2} \right\}^{1/\alpha} B_L \tag{18b}$$

Figure 5 shows the behavior of σ_ξ^2 and σ_n^2 as a function of loop bandwidth, as well as their sum, for $\alpha = 3$ and parameter values $S_1 = 4\pi \times 10^{-4}$ and $(N_0/2A^2) = 2 \times 10^{-2}$. This algorithm for loop bandwidth optimization requires only knowledge of σ_ξ^2, σ_n^2, and the bandwidth at which the measurements were made. Measurement of σ_ξ^2 also allows determination of the phase process constant S_1, via the equation

$$S_1 = \frac{\sigma_\xi^2\, B_L^{(\alpha-1)}}{2\,\gamma_0(\alpha)} \tag{19}$$

In a practical system, the required parameters and spectral densities are not known and hence must be estimated. The accuracy of the predictions ultimately depends on the accuracy of the estimates. The choice of best estimator structure often depends on the application. The selection of the estimator structure was guided by the fact that in our application the amount of available samples far exceeded the requirements imposed by resolution constraints.

E. Power Spectrum Estimator Using Bartlett's Procedure

The power spectral density estimates in this article use Bartlett's technique of averaged periodograms, which involves a trade-off between smoothing and spectral resolution. Bartlett's approach can be used to advantage whenever the available record length is so great that the required spectral resolution can be achieved with a small fraction of the available samples. Thus, if the total record length is K but $N \ll K$ samples satisfy the resolution requirements, then $M = K/N$ periodograms may be averaged to obtain a smoothed estimate of the spectral density. Our interest in periodograms stems from the fact that FFTs can be used to compute sampled versions of the desired periodograms quickly and efficiently.

The periodogram $I_N(\omega)$ associated with a sequence $y(n)$ of length N is defined as

$$I_N(\omega) = \frac{1}{N} |\, Y(e^{j\omega})\, |^2 \tag{20a}$$

where

$$Y(e^{j\omega}) = \sum_{n=0}^{N-1} y(n)\, e^{-j\omega n} \qquad \omega = 2\pi f \tag{20b}$$

is the discrete Fourier transform of the sequence $y(n)$. The use of periodograms in spectral estimation can be justified on the grounds that in the limit as N approaches infinity, the expected value of the periodogram approaches the power spectral density of $y(t)$, i.e.,

$$\lim_{N \to \infty} E\,[I_N(\omega)] = S_y(\omega) \tag{21}$$

Thus, in the limit the expected value of the periodogram is the desired power spectral density.

The variance of the periodogram as a function of frequency provides an indication of the amount of random variation about the mean of the estimate. For a real Gaussian sequence with power spectral density $S_y(\omega)$, the covariance is approximately [5]

$$\operatorname{cov}[I_N(\omega_1), I_N(\omega_2)] \cong S_y(\omega_1) S_y(\omega_2) \left\{ \left(\frac{\sin\left[(\omega_1 + \omega_2)\frac{N}{2}\right]}{N \sin\left[\frac{(\omega_1 - \omega_2)}{2}\right]} \right)^2 + \left(\frac{\sin\left[(\omega_1 - \omega_2)\frac{N}{2}\right]}{N \sin\left[\frac{(\omega_1 - \omega_2)}{2}\right]} \right)^2 \right\} \tag{22}$$

Note that the standard deviation of a single periodogram is at least as great as the spectral level itself, at any frequency. At frequencies away from zero, the standard deviation of $I_N(\omega)$ is well approximated by $I_N(\omega)$. Therefore, a single periodogram does not provide a very useful estimate of the power spectral density; hence averaging is usually required to reduce the variance of the spectral estimate.

The FFT can be employed to generate samples of $I_N(\omega)$. With

$$Y(k) \triangleq Y(e^{j\omega})\,|_{\omega = 2\pi k/N}$$

it follows that

$$I_N(k) = I_N(\omega)\,|_{\omega = 2\pi k/N}$$

Since the samples $y(n)$ were assumed to form a Gaussian sequence, it follows that the FFT outputs $Y(k)$ are Gaussian random variables. It is readily demonstrated that the random variables $Y(k)$ are uncorrelated and hence independent by the Gaussian assumption. It follows that the periodogram samples $I_N(k)$ are independent as well.

Let $\hat{S}_y(k)$ denote the final spectral estimate, after averaging over M periodograms. It is evident that

$$\lim_{N\to\infty} E[\hat{S}_y(k)] = \lim_{N\to\infty} \frac{1}{M} \sum_{j=1}^{M} E[I_{N,j}(k)] = S_y(k) \tag{23}$$

$$\operatorname{cov}[\hat{S}_y(k_1), \hat{S}_y(k_2)] \cong \frac{1}{M} S_y^2(k) \left[1 + \delta_{0,k_1}\right] \delta_{k_1,k_2} \tag{24}$$

where δ_{k_1,k_2} is the Kroenecker delta (this function is nonzero only for $k_1 = k_2$, where its value is one). In particular, the variance is

$$\operatorname{var}[\hat{S}_y(k)] = \frac{1}{M} S_y^2(k) \qquad k \neq 0 \tag{25}$$

The spectral estimate samples $\hat{S}_y(k)$ are independent random variables, by virtue of the fact that the final estimate is an average of M random variables. Thus for any k, averaging M independent periodograms reduces the variance of each estimate of spectral density by a factor of M relative to that of a single periodogram.

F. Error Analysis

The estimator described above cannot provide perfect estimates of the desired parameters by processing finite sequences (in fact, no estimator can). We determine the variance of the estimates $\hat{S}_1$ and $\hat{B}_L^*$, under the assumption that enough periodograms were averaged to ensure that the errors in the underlying parameters are but a small fraction of their actual values. Let $\hat{\sigma}_\xi^2$ and $\hat{\sigma}_n^2$ denote the estimates of σ_ξ^2 and σ_n^2, respectively. For small errors, we can write

$$\hat{\sigma}_\xi^2 \cong \sigma_\xi^2 + \epsilon_\xi^2 = \sigma_\xi^2 (1 + x_\xi) \tag{26a}$$

$$\hat{\sigma}_n^2 \cong \sigma_n^2 + \epsilon_n^2 = \sigma_n^2 (1 + x_n) \tag{26b}$$

where

$$x_\xi = \frac{\epsilon_\xi^2}{\sigma_\xi^2} \qquad x_n = \frac{\epsilon_n^2}{\sigma_n^2}$$

Assume that x_ξ and x_n are zero mean random variables, with rms values much less than one. Using Eqs. (18) and (19), the estimates of B_L^* and S_1 can be expressed in terms of their actual values as

$$\hat{B}_L^* = B_L^* \left\{ \frac{1 + x_\xi}{1 + x_n} \right\}^{1/\alpha} \tag{27a}$$

and

$$\hat{S}_1 = S_1 (1 + x_\xi) \tag{27b}$$

The second order statistics of S_1 can be obtained directly:

$$E[\hat{S}_1] = S_1 \tag{28a}$$

$$\text{var}[\hat{S}_1] = S_1^2 \,\text{var}(x_\zeta) = S_1^2 \frac{\text{var}(\hat{\sigma}_\zeta^2)}{\sigma_\zeta^4} \tag{28b}$$

(In fact, these second order statistics for S_1 are valid in general.) To obtain the statistics of $\hat{B}_L^*$, expand the ratio containing the error terms and ignore all but the linear terms:

$$\left\{\frac{1+x_\zeta}{1+x_n}\right\}^{1/\alpha} = \left\{(1+x_\zeta)\left(1-x_n+\frac{x_n^2}{2}-\cdots\right)\right\}^{1/\alpha}$$

$$\cong \left\{1+(x_\zeta - x_n)\right\}^{1/\alpha}$$

$$\cong 1+\frac{1}{\alpha}(x_\zeta - x_n) \tag{29}$$

Thus, for suitably small errors, we obtain the approximation

$$\hat{B}_L^* \cong B_L^* \left\{1+\frac{1}{\alpha}(x_\zeta - x_n)\right\} \tag{30}$$

It follows that

$$E[\hat{B}_L^*] \cong B_L^* \tag{31a}$$

$$\text{var}[\hat{B}_L^*] = \left(\frac{B_L^*}{\alpha}\right)^2 \text{var}(x_\zeta - x_n)$$

$$\leqslant \left(\frac{B_L^*}{\alpha}\right)^2 \left\{\frac{\text{var}(\hat{\sigma}_\zeta^2)}{\sigma_\zeta^4} + \frac{\text{var}(\hat{\sigma}_n^2)}{\sigma_n^4} + 2\frac{\sqrt{\text{var}(\hat{\sigma}_\zeta^2)}\sqrt{\text{var}(\hat{\sigma}_n^2)}}{\sigma_\zeta^2\,\sigma_n^2}\right\} \tag{31b}$$

The upper bound in Eq. (31b) is useful when the correlation between x_ζ and x_n is not well known. Actual numerical values will be determined in the following section, where the techniques for obtaining the parameters from the spectral estimates are examined in greater detail.

III. Pioneer 10 Results

This section presents the results of experiments in tracking the Pioneer 10 spacecraft with the breadboard DSN Advanced Receiver [6]. The data were obtained using the 64 meter antenna at DSS-14 in Goldstone, California. The data include both one-way and three-way transmission modes obtained at various SEP angles, illustrating the effect of the solar corona on carrier phase. Spectra were measured and loop bandwidths were optimized for the various conditions.

The Advanced Receiver is a hybrid analog/digital receiver which makes digital estimates of the carrier phase and filters these estimates with a digital second or third order tracking filter. The filter output drives an analog frequency synthesizer, closing the loop. The phase detector output provides the datastream which is subsequently analyzed to determine the received spectral densities. The power spectral density at the phase detector output, S_y, was obtained as described in Section II.E. These estimates were obtained in near-real time. Phase measurements as a function of time were clocked out of the Advanced Receiver to a control computer, where these measurements were recorded on floppy discs. After sufficient data were recorded, the disc was processed on an IBM PC. The processing was typically completed in less than five minutes, during which time data for the next tracking case was collected by the Advanced Receiver.

A. Spectral Estimation Results

Estimates of the power spectral density of $y(n)$, $\hat{S}_y(k)$, were obtained by means of averaged periodograms, using the FFT. If the sampling rate satisfies the Nyquist criterion, the result is also a good estimate of the spectrum of the continuous process $y(t)$. Results for one-way transmission are shown in Figs. 6a, b, and c. The phase spectral density estimates are represented by the solid curve (the points of the FFT output are connected by straight lines). For the first case, Fig. 6a, the number of points in the sequence was $N = 128$ and $M = 25$ records were averaged. The sampling rate was 8 samples per second. Since for real samples the positive and negative portions of the spectrum are identical, only the positive portion is displayed. For purposes of interpolation, 128 zeros were appended to each record prior to processing, and a 256 point FFT performed (this does not change the total number of independent samples in the FFT output, which remains N). In Fig. 6, a third order loop with parameters $r = 4$, $k = 0.25$ was used, with bandwidths of 0.5 Hz, 0.8 Hz, and 2.0 Hz (the 0.8 Hz bandwidth is nearly optimum). The contribution of the process spectrum near the origin dominates for the 0.5 Hz loop of Fig. 6a; it is still present in the 0.8 Hz loop (Fig. 6b), but is not discernible for the 2.0 Hz loop (Fig. 6c). The dominance of the white noise component is apparent in all cases at frequencies greater than the loop bandwidth.

The above observation suggests the following procedure for estimating the desired components of the spectral density:

(1) Estimate the normalized spectral level $N_0/2A^2$ by averaging the discrete points of the spectral estimate over the upper half of the spectrum. (Henceforth, cap and Est[•] shall both denote "estimate.")

(2) The function $|1 - H(jk)|^2 S_n(k)$ is estimated, assuming the filter transfer function is known.

(3) An estimate of the filtered process spectrum is obtained from Eq. (5b) as

$$\text{Est}\left\{|1 - H(jk)|^2 S_\xi(k)\right\} = \frac{1}{A^2}\left\{\hat{S}_y(k) - |1 - H(jk)|^2 \hat{S}_n(k)\right\} \tag{32}$$

(4) Finally, having decomposed the estimated observation spectral density into its components, estimates of the underlying spectral densities are obtained.

Because the signal amplitude A is assumed known, it follows that

$$\hat{S}_n(k) = A^2\left(\frac{\hat{S}_n(k)}{A^2}\right) \tag{33}$$

Since a $2N$-point FFT was performed on an N-point sequence, the resulting $N/2$-point average has $N/4$ degrees of freedom and hence may be treated as the average of $N/4$ independent random variables. Taking into account the M records that were averaged to obtain the spectral density estimate, the mean and variance of this estimate are

$$E\left[\frac{\hat{S}_n(k)}{A^2}\right] = \frac{N_0}{2A^2} \tag{34}$$

and

$$\text{var}\left[\frac{\hat{S}_n(k)}{A^2}\right] = \frac{4}{NM}\left(\frac{N_0}{2A^2}\right)^2 \tag{35}$$

The smooth dashed curves in Figs. 6a through 6c are the estimates of receiver noise spectra filtered by $(1 - H(jk))$, while the rough dashed curves are the estimates of the filtered phase process spectral density.

The same techniques can be applied to the case of three-way telemetry. Results for three-way tracking on DOY 141 are shown in Fig. 7, again around the optimum loop bandwidth, which in this case was roughly 0.6 Hz. For this day, the SEP angle was 11.8 degrees. This resulted in process noise due to solar scintillation which was almost as high as the one-way phase noise due to the spacecraft oscillator. Estimates of the various system and model parameters can be obtained from these fundamental estimates. The following examples serve to illustrate some techniques for deriving the desired parameter estimates.

B. Phase Error Variance Estimates

The components of the phase error variance can be found from the measured spectra. If the loop bandwidth is known, then from Eq. (15b) the variance of the additive noise component can be found since it is proportional to the normalized noise spectral level estimate:

$$\hat{\sigma}_n^2 = 2B_L\left(\frac{\hat{N}_0}{2A^2}\right) \tag{36a}$$

Using the estimated spectral level, the variance of $\hat{\sigma}_n^2$ is approximately

$$\text{var}(\hat{\sigma}_n^2) \cong \frac{4}{NM}(2B_L)^2\left(\frac{\hat{N}_0}{2A^2}\right)^2 = \frac{4}{NM}(\hat{\sigma}_n^2)^2 \tag{36b}$$

Computation of the variance due to the phase process requires integrating the process spectral estimate. For suitably great N, integration over the lower half of the positive spectrum can be approximated by the sum of independent random variables as

$$\hat{\sigma}_\xi^2 = 2\int_0^\infty |1 - H(j2\pi f)|^2 \hat{S}_\xi(f)\,df$$

$$\cong 2\left\{\Delta f \sum_{k=0,2,4,\cdots}^{\frac{N}{2}-2} |1 - H(jk)|^2 \hat{S}_\xi(k)\right\} \tag{37a}$$

where F is the total frequency range and $\Delta f = F/N$ is the effective spectral resolution of the FFT. Using Eqs. (32) and (35) in Eq. (37a), and since the spectral level estimates obtained from the upper half of the spectrum are independent of the lower half, the variance of the estimate can be expressed as

$$\text{var}(\hat{\sigma}_\xi^2) \cong 4(\Delta f)^2 \left\{ \sum_{k=0,2,4,\cdots}^{\frac{N}{2}-2} \text{var}\left(\frac{\hat{S}_y(k)}{A^2}\right) \right\}$$

$$+ 4(\Delta f)^2 \left(\frac{4}{NM}\right)\left(\frac{\hat{N}_0}{2A^2}\right)^2 \left\{ \sum_{k=0,2,4,\cdots}^{\frac{N}{2}-2} |1 - H(jk)|^2 \right\}^2 \tag{37b}$$

where we used the fact that the $N/2$-point estimate has only $N/4$ degrees of freedom. Near the optimum loop bandwidth the observed spectral level is essentially constant and equal to the normalized spectral level of the additive noise (see Figs. 6 and 7). Upper bounding $|1 - H(jk)|^2$ by one, Eq. (37b) can be approximated as

$$\text{var}(\hat{\sigma}_\xi^2) \cong \frac{4(\Delta f)^2}{M}\left(\frac{N}{4}+1\right)\left(\frac{\hat{N}_0}{2A^2}\right)^2$$

$$= \frac{4(\Delta f)^2}{M}\left(\frac{N}{4}+1\right)\left(\frac{\hat{\sigma}_n^2}{2B_L}\right)^2 \tag{37c}$$

For the Pioneer 10 data obtained on DOY 119 and 141, the total phase error, i.e., the sum of the two components, is plotted as functions of loop bandwidth in Fig. 8, along with error bars of one standard deviation. Both one-way and three-way results are shown. It is apparent that in this case three-way reception is dominated by solar scintillation effects.

C. Optimum Loop Bandwidth

Next we consider the estimation of the optimum loop bandwidth by means of Eq. (18):

$$\hat{B}_L^* = \left[\frac{(\alpha-1)\,\hat{\sigma}_\xi^2}{\hat{\sigma}_n^2}\right]^{1/\alpha} B_L \tag{38}$$

The estimate of the optimum loop bandwidth depends on the ratio of phase process and white noise variances, as well as on the loop bandwidth at which these variances are measured. Table 1 presents the estimated parameters for all data taken during Pioneer 10 tracking. At least three different loop bandwidths were used each day.

The estimates of optimum loop bandwidth are consistent on each day, except for the cases in Table 1 that are flagged by a question mark. These estimates disagree with the other estimates by significantly more than the estimation error. These estimates depend on $\hat{\sigma}_\xi^2$, which is the difference of two relatively large numbers whose difference is small. Thus the actual errors may be dominated by small biases that were not accounted for in the estimation of the larger quantities, one of which is $(N_0/2A^2)$.

In Table 1, the estimate

$$\frac{P_c}{N_0} = \left[2\left(\frac{\hat{N}_0}{2A^2}\right)\right]^{-1} \tag{39}$$

is displayed (in dB-Hz) instead of the normalized noise spectral level, in order to facilitate comparison with standard experimental measurements.

The variances of parameter estimates are displayed in Table 2. These variances were computed using the constant spectral level assumption, and using the parameter estimates instead of their true (but unknown) values in Eqs. (28b) and (31b). Because $\hat{\sigma}_n^2$ and $\hat{\sigma}_\xi^2$ are not highly correlated, we use the approximation $\text{var}(\hat{\sigma}_\phi^2) \cong \text{var}(\hat{\sigma}_n^2) + \text{var}(\hat{\sigma}_\xi^2)$. The variances of $\hat{B}_L^*$ were computed by means of the upper bound in Eq. (31b) and hence tend to be too large. The standard deviation error bars in Figs. 8 and 9 were obtained directly from Table 2. For bandwidths smaller than the optimum, the actual standard deviations may be larger than indicated, because the spectral peak tends to be much greater than the spectral level of the additive noise used in the calculations.

D. Phase Process Spectral Estimates

Estimates of the phase process spectral density at unit frequency S_1 are presented for all of the Pioneer data in Fig. 9 and in Table 1. The spectral density at 1 Hz, S_1, was estimated by means of Eq. (19). The f^{-3} model was used for all one-way data, and the $f^{-8/3}$ model was used for all three-way data. The SEP angle is indicated in Fig. 9 for each day data was collected. The results show that one-way data have spectral levels of $S_1 = 1$ to 2×10^{-3} r^2/Hz.

For the three-way data, the results depend on SEP angle, indicating that here solar scintillation effects dominate. The measured spectral levels were approximately 10^{-4} r^2/Hz on DOY 128, when the SEP angle was 23.7 deg, and 10^{-3} r^2/Hz on DOY 141 at an SEP angle of 11.8 deg. Armstrong, Woo, and Estabrook [1] have measured solar scintillation dominated carrier spectra using the Viking spacecraft. At $f = 0.001$ Hz, they observed a spectral level of 10^3 to 10^5 r^2/Hz at 23.7 deg

and 10^4 to 10^6 r^2/Hz at 11.8 deg for one-way paths at a carrier frequency of 2.3 GHz. Multiplying by 10^8 to convert to S_1 using the $\alpha = 8/3$ assumption, and also multiplying by 2 to convert to three-way reception, the corresponding range for S_1 becomes 2×10^{-5} to 2×10^{-3} at 23.7 deg and 2×10^{-4} to 2×10^{-2} at 11.8 deg. Although extrapolation to frequencies as high as 1 Hz may not be accurate, we observe that our results are consistent with the previously measured values. The variation in the S_1 estimates evident in the three-way mode (Fig. 9) could be the result of actual solar scintillation, or could be due to modeling errors. For example, during three-way reception, the phase noise is due to a combination of spacecraft transmitter phase effects and solar scintillation; hence the effective exponent α may not be exactly 8/3, as was assumed in the calculations. In addition, the loop bandwidth in the Advanced Receiver is not precisely known. The estimates of S_1 are sensitive to both of these effects. However, since loop bandwidth optimization is less critically dependent on B_L and α, our model is sufficiently accurate for the purpose of optimizing receiver loop bandwidth. Since the spectra measured in [1] tend to exhibit significant fluctuations, our method of measuring spectra in real time should prove to be very useful for optimizing loop bandwidths and for obtaining other useful system parameter estimates during telemetry.

IV. Conclusions

A method for optimizing loop bandwidths via spectral estimation for phase locked loop receivers has been developed and used to improve carrier tracking for the Pioneer 10 spacecraft. Estimates were made of the relevant spectral parameters and of the total noise power for one-way and three-way transmission, and the variance of the phase error in the loop was minimized. Results obtained in the field were found to be in good agreement with theoretical values. In addition, by analyzing phase spectra, solar channel effects were detected and compensated for. Thus the Fourier analysis of the phase detector output was shown to be an important tool which in addition to bandwidth optimization can be used to monitor channel effects, transmitted carrier stability, and make estimates of relevant system parameters.

It was specifically shown for Pioneer 10 that receiver loop bandwidths of 0.1 to 1.0 Hz are feasible, depending on the type of transmission (one-way or three-way), and depending on the SEP angle. When compared to the 3 Hz loop bandwidths used with operational DSN receivers, it is clear that the narrower loop bandwidths achieve improvements of 5 to 15 dB in carrier tracking threshold and carrier loop SNR.

References

[1] J. W. Armstrong, R. Woo, and F. B. Estabrook, "Interplanetary Phase Scintillation and the Search for Very Low Frequency Gravitational Radiation," *The Astrophysical Journal*, vol. 230, pp. 570-574, June 1, 1979.

[2] S. Aguirre and W. J. Hurd, "Design and Performance of Sampled Data Loops for Subcarrier and Carrier Tracking," *TDA Progress Report 42-79*, vol. July-September 1984, Jet Propulsion Laboratory, Pasadena, California, pp. 81-95, November 15, 1984.

[3] R. Gagliardi, *Introduction to Communications Engineering*, New York: John Wiley & Sons, 1978.

[4] D. Halford, "A General Mechanical Model for $|f|^{\alpha}$ Spectral Density Random Noise with Special Reference to Flicker Noise $1/|f|$," *Proc. IEEE*, vol. 56, pp. 251-258, March 1986.

[5] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, New York: Prentice-Hall, 1975.

[6] D. H. Brown and W. J. Hurd, "DSN Advanced Receiver: Breadboard Description and Test Results," *TDA Progress Report 42-89*, vol. January-March 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 48-66, May 15, 1987.

Table 1. Parameter estimates

DOY	GMT	B_L (Hz)	SEP	$\left(\frac{\widehat{P_c}}{N_0}\right)$ (dB–Hz)	$\widehat{\sigma_\xi^2}$ ($\times 10^{-2}$)	$\widehat{\sigma_n^2}$ ($\times 10^{-2}$)	$\widehat{\sigma_\phi^2}$ ($\times 10^{-2}$)	$\widehat{B_L^*}$ (Hz)	$\widehat{S}_1$ ($\times 10^{-4}$)	
119	2:39	0.5		11.6	7.79	3.44	11.23	0.83	10.7	
119	2:02	0.7	32.4°	12.0	4.42	4.42	8.84	0.88	11.9	
119	3:32	0.8		10.6	3.17	6.91	10.08	0.78	11.2	
119	4:03	1.0		9.8	3.91	10.4	14.31	0.91	21.5 ?	One Way
140	22:14	0.5		13.2	8.76	2.41	11.17	0.97	12.1	
140	22:46	0.8	11.8°	13.5	3.07	3.60	6.67	0.96	10.8	
140	22:30	2.0		12.2	2.07	12.2	14.27	1.40	45.6 ?	
128	3:04	0.125		12.6	5.25	0.69	5.94	0.32	1.40	
128	3:25	0.25	23.7°	11.8	1.29	1.65	2.94	0.28	1.09	
128	2:42	0.5		12.6	0.81	2.74	3.55	0.38	2.2	
141	1:30	0.25		12.3	8.55	1.46	10.01	0.59	7.2	Three Way
141	1:55	0.375		11.9	5.43	2.40	7.83	0.62	9.0	
141	0:51	0.5	11.8°	12.6	3.28	2.75	6.03	0.65	8.8	
141	0.12	0.75		12.9	2.33	3.87	6.20	0.75	12.3	
140	23:32	1.0		12.7	1.75	5.44	7.19	0.79	14.9	

Table 2. Variances of parameter estimates

DOY	B_L (Hz)	$\text{var}\left(\frac{\widehat{N_0}}{2A^2}\right)$ ($\times 10^{-7}$)	$\text{var}(\widehat{\sigma_\phi^2})$ ($\times 10^{-6}$)	$\text{var}(\widehat{S}_1)$ ($\times 10^{-9}$)	$\text{var}(\widehat{B_L^*})$ ($\times 10^{-4}$)	
119	0.5	5.1	8.9	1.6	2.6	
119	0.7	4.3	7.9	5.1	5.7	
119	0.8	16	31	33	25	
119	1.0	34	69	170	47	One Way
140	0.5	7.3	13	2.3	5.8	
140	0.8	4.4	8.4	9	14	
140	2.0	8.1	26	650	92	
128	0.125	7.4	3.1	0.02	0.61	
128	0.25	5.2	2.3	0.15	2.0	
128	0.5	19	9.9	5.7	33	
141	0.25	7.4	3.2	0.22	1.2	Three Way
141	0.375	8.9	4.2	1.0	2.2	
141	0.5	3.9	2.0	1.2	2.2	
141	0.75	2.1	3.9	9.5	7.5	
140	1.0	2.3	4.7	28	15	

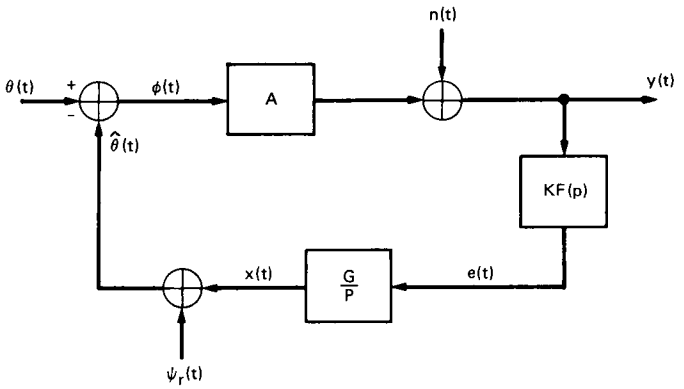


Fig. 1. Receiver phase locked loop block diagram

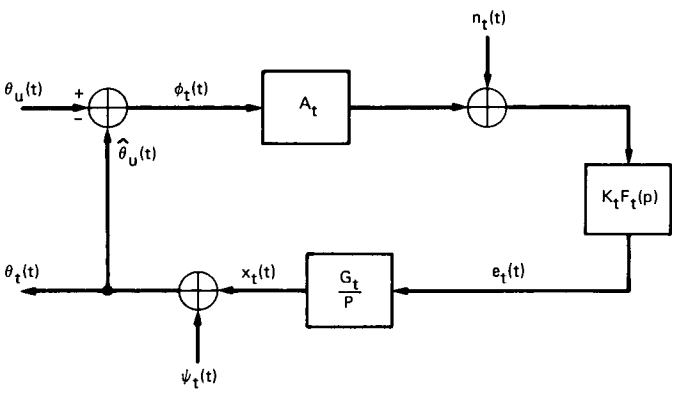


Fig. 2. Spacecraft transmitter phase locked loop block diagram

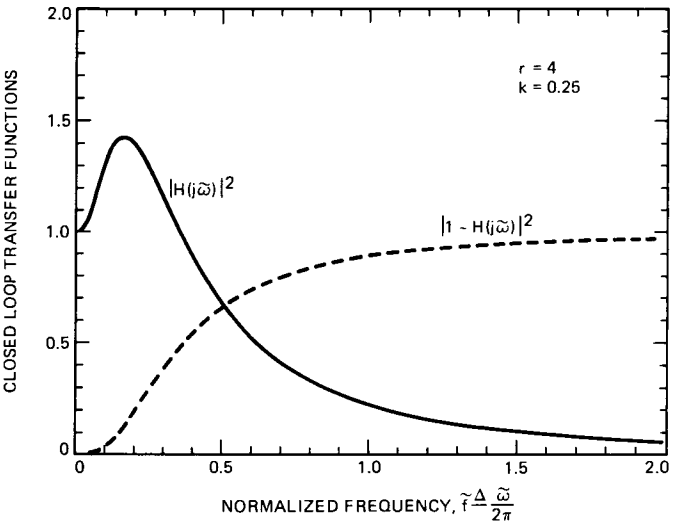


Fig. 3. Normalized closed loop transfer functions

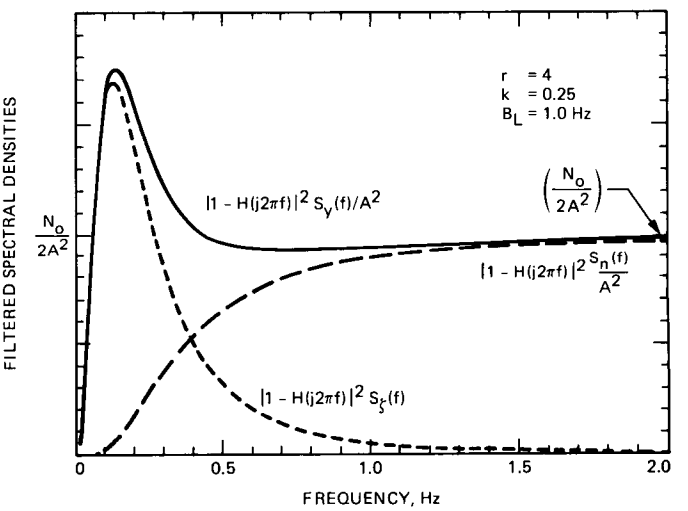


Fig. 4. Filtered spectral densities for a third order loop (inverse f^3 phase process and additive white noise)

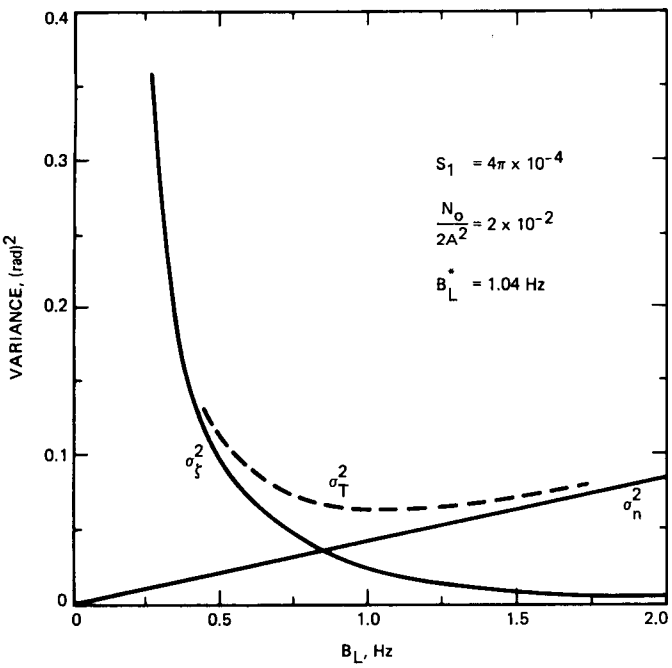


Fig. 5. Phase error variance and components as a function of loop bandwidth ($\alpha = 3$)

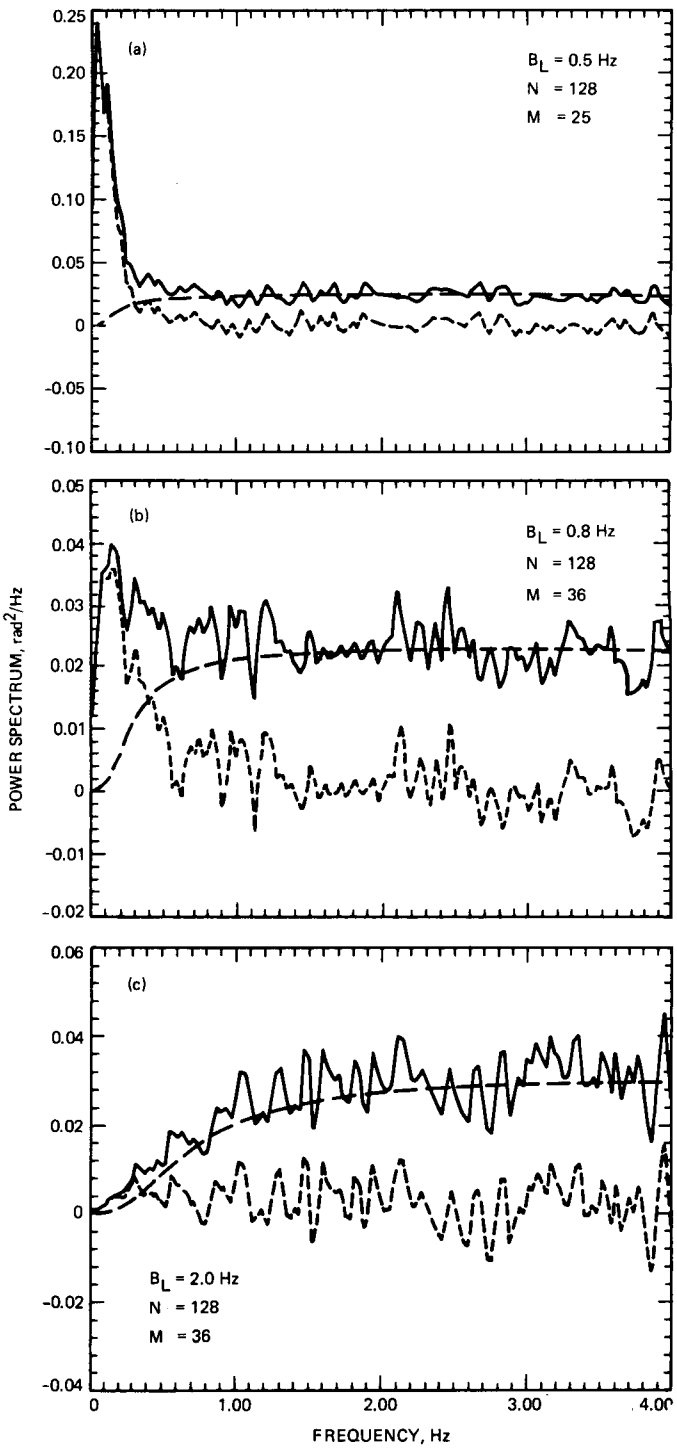


Fig. 6. Spectral density estimates (one-way mode)

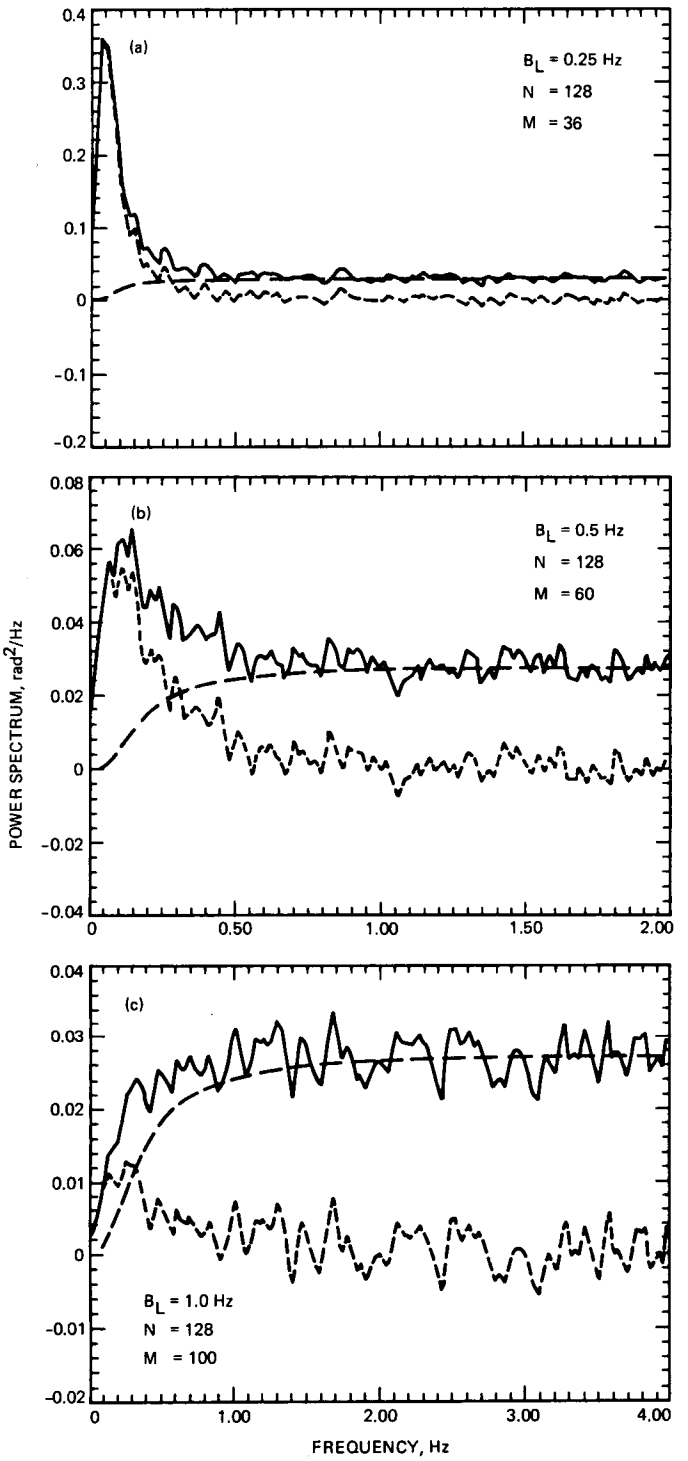


Fig. 7. Spectral density estimates (three-way mode)

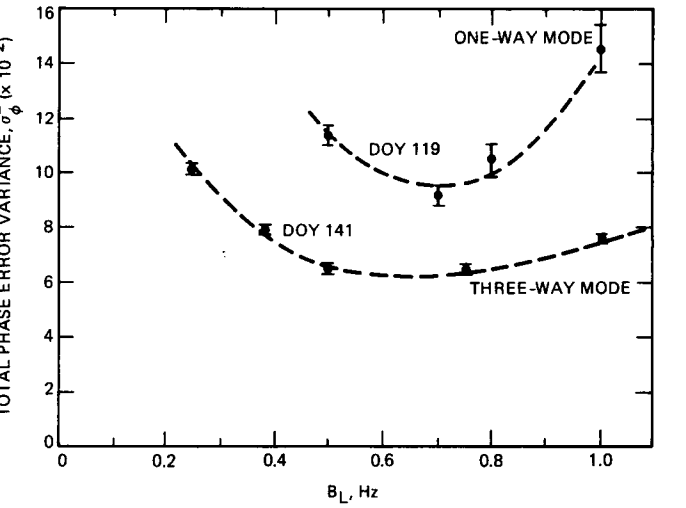


Fig. 8. Measured total phase error variances as a function of loop bandwidth

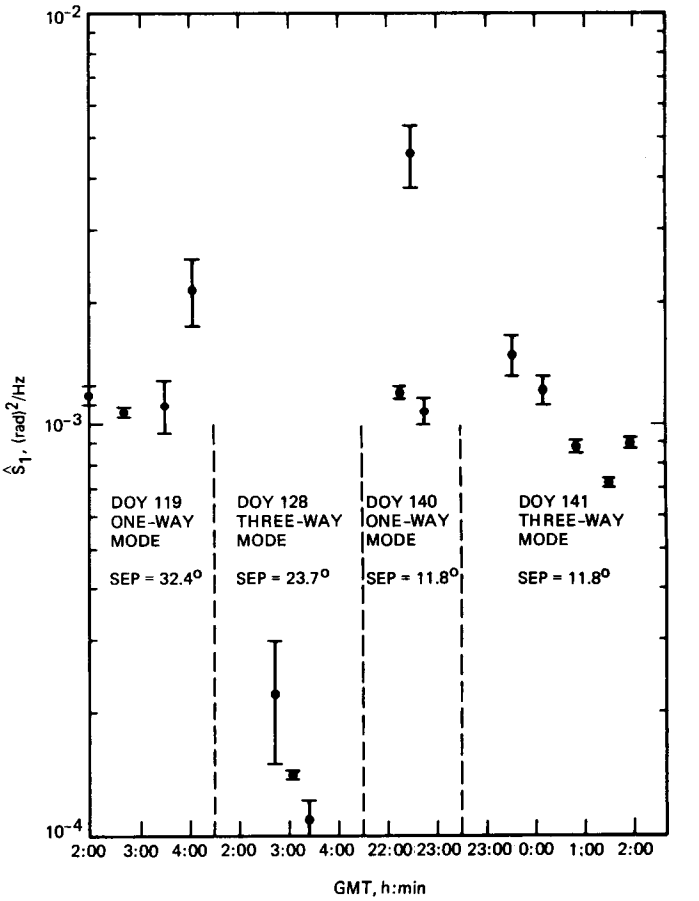


Fig. 9. Estimates of phase spectrum parameter

Appendix

Maximum of Filtered Spectrum

Here we demonstrate that for f^{-3} type phase processes the maximum of the filtered process spectrum $|1 - H(j\,2\pi f)|^2 S_\xi(f)$ is always at a given fraction of the loop bandwidth, for transfer functions of the form defined in Eq. (7) and inverse f^3 process spectrum defined in Eq. (11). These functions can be expressed in terms of the radian frequency $\omega = 2\pi f$ as

$$S_\xi(\omega) = (2\pi)^3 \frac{S_1}{|\omega|^3} \tag{A1}$$

and

$$|1 - H(j\omega)|^2 S_\xi(\omega) = (2\pi)^3 S_1 G(\omega; r, k, \tau_2) \tag{A2}$$

where

$$G \triangleq \omega^3 [\omega^6 + \alpha_1 \omega^4 + \alpha_2 \omega^2 + \alpha_3]^{-1}$$

$$\alpha_1 = \frac{r(r-1)}{\tau_2^2} \qquad \alpha_2 = \frac{r^2(1-2k)}{\tau_2^4} \tag{A3}$$

$$\alpha_3 = \frac{r^2 k^2}{\tau_2^6}$$

Setting $y = \omega^2$, differentiating Eq. (A3) with respect to y, and setting the result equal to zero yields the cubic equation

$$y^3 + \left(\frac{\alpha_1}{3}\right) y^2 - \left(\frac{\alpha_2}{3}\right) y - \alpha_3 = 0 \tag{A4}$$

Further, letting $x = y + (\alpha_1/9)$ results in the simplified form

$$x^3 + ax + b = 0 \tag{A5}$$

where

$$a = \frac{1}{3}\left(-\alpha_2 - \left(\frac{\alpha_1^2}{9}\right)\right) \tag{A6a}$$

$$b = \frac{1}{27}\left(2\frac{\alpha_1^3}{27} + \alpha_1\alpha_2 - 27\alpha_3\right) \tag{A6b}$$

Trigonometric solutions for x are obtained by letting $x = m\cos\theta$, with $m = 2\sqrt{-a/3}$. The solutions for y are of the form

$$y_i = (2\pi)^2 C_i^2 (r, k) B_L^2 \tag{A7}$$

Taking only the positive solutions, we obtain

$$\omega^* = 2\pi C_0(r, k) B_L \tag{A8}$$

Direct evaluation yields the location of the maxima for $r = 4$

$$k = 0: \; f^* = C_0(4, 0) B_L = 0.147 B_L \tag{A9a}$$

$$k = 0.25: \; f^* = C_0(4, 0.25) B_L = 0.127 B_L \tag{A9b}$$

These results are confirmed by the graphs shown in Fig. A1, where the location of the maxima are seen to be at the predicted frequencies. This property helps to determine if the observed phase process is indeed an f^{-3} type process.

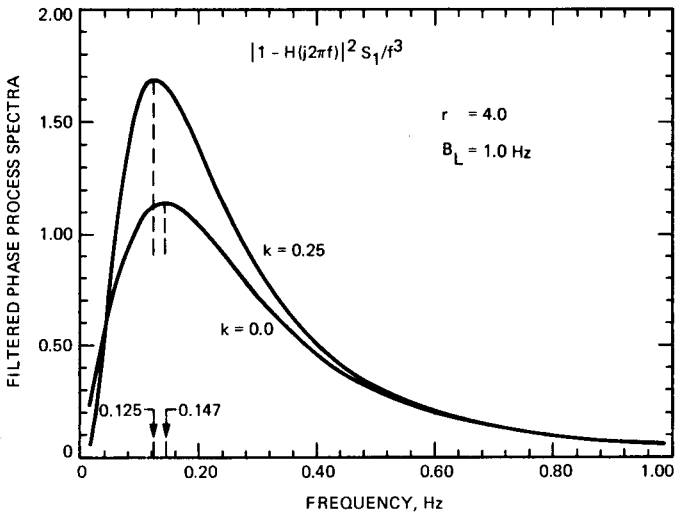


Fig. A1. Filtered process spectra for second and third order loops

Detection of Signals by the Digital Integrate-and-Dump Filter With Offset Sampling

R. Sadr and W. J. Hurd
Communications Systems Research Section

The Integrate-and-Dump Filter (IDF) is used as a matched filter for the detection of signals in additive white Gaussian noise. In this article, the performance of the digital integrate-and-dump filter is evaluated. The case considered is when symbol times are known and the sampling clock is free running at a constant rate, i.e., the sampling clock is not phase locked to the symbol clock. Degradations in the output signal-to-noise ratio of the digital implementation due to sampling rate, sampling offset, and finite bandwidth, resulting from the anti-aliasing low-pass prefilter, are computed and compared with those of the analog counterpart. It is shown that the digital IDF performs within 0.6 dB of the ideal analog IDF whenever the prefilter bandwidth exceeds four times the symbol rate and when sampling is performed at the Nyquist rate. The loss can be reduced to 0.3 dB by doubling the sampling rate, where 0.2 dB loss results from finite bandwidth and 0.1 dB results from the digital IDF.

I. Introduction

An Integrate-and-Dump Filter (IDF) is the ideal matched filter for coherent detection of rectangular pulse shape signals corrupted by additive white Gaussian noise (AWGN). A digital implementation of the IDF has numerous advantages over its analog counterpart, such as the ability to dump instantaneously with no overshoot, freedom of drift from the quiescent operating point, and the use of advanced digital integrated circuits (ICs) to perform multiplication and accumulation with greater accuracy than the analog counterparts.

A. Current Problem

The digital implementation of the IDF requires the input waveform to be sampled. This requires that the input signal be filtered prior to sampling to eliminate aliasing. Throughout this article, the anti-aliasing low-pass prefilter is also referred to simply as the filter. Filtering of the observed signal results in the transmitted signal pulses becoming band limited and also causes Inter-Symbol Interference (ISI). The finite bandwidth of the observed signal and ISI both degrade the performance of the receiver. In order to evaluate the performance versus the filter bandwidth and sampling frequency, the analog IDF depicted in Fig. 1(a) is compared with the performance of the digital IDF shown in Fig. 1(b).

In the limit, as the prefilter low-pass bandwidth W Hz and, as a consequence, the sampling rate approach infinity, the performance of the digital IDF converges to that of the analog IDF. However, this is not practical from an implementation

point of view, since large bandwidths require higher sampling rates to satisfy the Nyquist criterion. In this article, we find the performance degradation which results from filtering the received signal and from using only a small finite number of samples.

Furthermore, we also superimpose the effect of "offset sampling" in the performance of the digital IDF. When sampling a signal of finite duration T sec with a sampling period of T_s sec, the first sample of the signal may occur anywhere in the time interval $0 \leqslant t < T_s$. This consideration for the digital IDF has not been studied previously. In this article, in addition to analyzing the effects of prefiltering and sampling on performance, we seek to find the degradation due to this offset.

In the advanced receiver being developed for the NASA Deep Space Network, a wide range of symbol rates must be processed with the maximum available bandwidth and sampling rate. When the number of samples per symbol is large, the loss due to the offset in sampling is negligible. This loss is not negligible at high symbol rates, when the number of samples per symbol is not large.

In the digital IDF implementation, we assume that the sampling clock is free running at a constant rate, i.e., that the sampling clock is not phase locked to a multiple of the symbol clock. We also assume perfect symbol synchronization in the sense that the receiver has perfect information regarding the time that each symbol starts and ends.

For all of the results presented in this article except Section V.E, the symbol period is an integer multiple of the sampling period, and the performance is determined as a function of the relative phase offset between the sample times and the symbol times. In Section V.E, we consider the case for which the symbol period is not an integer multiple of the sampling period. We ignore quantization noise and oscillator instability in this article.

B. Previous Work

The digital IDF was studied previously by Natali [1] in 1969. In [1], a first order low-pass filter was considered and the signal-to-noise ratio (SNR) loss was analyzed for different time bandwidth products. Only a single symbol was considered. Our results match those of Natali for the case of no offset, a single symbol, and using a first order low-pass filter.

In 1969, Hartman [2] studied the degradation of SNR due to intermediate frequency (IF) filtering. He computed the SNR degradation due to bandwidth limiting of a biphase modulated signal when using an analog IDF. A lower boundary for the SNR was derived for different time-bandwidth products.

In 1976, Turin [3] analyzed the noncoherent digital matched filter matched to amplitude-modulated (AM) signals in the presence of additive white Gaussian noise and jamming. He showed that in the presence of jamming, improvement is possible over the analog matched filter by threshold biasing and dithering techniques for demodulation.

In 1978, Lim [4] analyzed the noncoherent digital matched filters with multibit quantization, matched to phase-shift-keyed signals.

In 1979, Chie [5] analyzed the digital IDF filter. The performance of the digital IDF was evaluated in relation to the number of bits used by the analog-to-digital (A/D) converter, the bandwidth of the prefilter, and the gain loading factor of the A/D converter. He also considered quantization error and the accumulator length. To study the performance of the digital IDF, he considered the symbol error probability resulting from hard quantization of the output of the digital IDF.

It is difficult to determine the exact symbol error probability in the presence of ISI, which is inherent in digital IDF systems. The ISI is caused by the low-pass filtering of the input signal. Recently, Helstrom [6] and Levy [7] outlined general algorithms for approximation of the symbol error probability. We should note here that numerous articles and techniques address the approximation of the symbol error probability in the presence of ISI. Citing all the references on this subject is beyond the scope of this article. Furthermore, our intended application for this article is the advanced receiver for NASA's Deep Space Network. The receiver output detected symbol values are quantized to several bits of accuracy (soft decision), as opposed to making hard decisions on the symbols. These quantized symbols are then used by the decoder for estimating the transmitted bit sequence. In our case, the SNR is therefore a more relevant parameter than the symbol error probability.

We make the fundamental assumption that the sampling clock is stable and the receiver symbol clock is synchronized. The effects of transmitter receiver clock time-base instability on coherent communication systems was analyzed by Chie [8] in 1982. The types of time-base instability modeled and analyzed are bit jitter and bit jitter rate.

C. Outline of Article

In Section II, the IDF is formulated and the average signal response and noise variance are derived. In Section III, the expression for the loss is formulated. In Section IV, the loss is expressed for two special cases: the first order low-pass filter and the ideal low-pass filter. In Section V, the numerical results are presented for known waveforms and for pseudo-random data patterns generated by Monte Carlo simulation.

Selections of bandwidth, sampling rate, and asynchronous sampling are discussed in this section. This work is summarized and conclusions are drawn in Section VI, and Section VII suggests a direction for future work.

II. System Description

The received signal plus noise is denoted by $r(t) = s(t - \tau_0) + n(t)$, where $s(t)$ is the signal, $n(t)$ is the noise, and τ_0 is the delay from the transmitter to the receiver. The transmitted signal $s(t)$ is a sequence of pulses expressed as

$$s(t) = \sum_k a_k\, p(t - kT) \tag{1}$$

At this point we impose no restriction on the shape or duration of each pulse $p(t)$. The input alphabet U is a finite alphabet with $a_i \in U = \{\pm 1, \pm 2, \ldots, \pm M\}$.

The analog IDF is shown in Fig. 1(a). The analog system is an ideal matched filter when $p(t)$ is a rectangular pulse from $t = 0$ to $t = T$. It detects the kth symbol by integrating over time $kT + \tau_0$ to $(k + 1)T + \tau_0$. The digital IDF is depicted in Fig. 1(b). In the digital implementation, an anti-aliasing low-pass prefilter is used for filtering the input signal. The filter output is sampled, with the ith sample occurring at time $iT_s + \tau_1$. The digital IDF detects the kth symbol by summing all the samples from $t = kT + \tau_0$ to $t = (k + 1)T + \tau_0$.

We assume that there is perfect symbol synchronization at the receiver, so the beginning and end times of each symbol are known. For the kth symbol, the "sampling offset" is defined by the time difference between the start of the symbol and the first sample within the symbol time, i.e., $(iT_s + \tau_1) - (kT + \tau_0)$. The first sample within each symbol time may occur anywhere between 0 and T_s seconds. A typical symbol waveform and the sampling points are shown in Fig. 2. We seek to determine the signal-to-noise ratio (SNR) for the output sample of this system.

Initially, we consider the average signal response of the digital IDF; later we consider the noise response.

A. Average Signal Response

The response of the low-pass filter to the observed signal $r(t)$ is

$$y(t) = \int_{-\infty}^{\infty} h(t - \zeta)\, s(\zeta - \tau_0)\, d\zeta + \int_{-\infty}^{\infty} h(t - \zeta)\, n(\zeta)\, d\zeta \tag{2}$$

Using Eq. (1) for $s(t)$ we have

$$y(t) = \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} a_k h(t - \zeta)\, p(\zeta - kT - \tau_0)\, d\zeta + \int_{-\infty}^{\infty} h(t - \zeta)\, n(\zeta)\, d\zeta \tag{3}$$

This signal $y(t)$ is sampled each T_s sec at time $t = iT_s + \tau_1$. We denote $y(iT_s + \tau_1)$ as y_i. Taking the expectation of Eq. (3) conditioned on a given data sequence a, and noting that the noise $n(t)$ is assumed to have zero mean, the conditional expectation of y_i is

$$E[y_i \mid \underline{a}] = \sum_{k=-\infty}^{\infty} a_k \int_{-\infty}^{\infty} h(iT_s + \tau_1 - \zeta)\, p(\zeta - kT - \tau_0) d\zeta \tag{4}$$

With a change of variable, Eq. (4) can be written as

$$E[y_i \mid \underline{a}] = \sum_k a_k \int_{-\infty}^{\infty} h(iT_s - kT + \delta - x)\, p(x) dx \tag{5}$$

where $\delta = \tau_1 - \tau_0$. Let

$$q_i(k, \delta) = \int_{-\infty}^{\infty} h(iT_s - kT + \delta - x)\, p(x) dx \tag{6}$$

represent the signal response of the filter at time $iT_s + \tau_1$ due to a single pulse at time $kT + \tau_0$. For simplicity we denote $q_i(k, \delta)$, as $q_i(k)$. The total average signal response from Eq. (5) for a given fixed δ may be expressed as

$$E[y_i \mid \underline{a}, \delta] = \sum_k a_k\, q_i(k) \tag{7}$$

Let I^k be the set of all i such that the ith sample falls in the kth symbol time, i.e.,

$$I^k = \{i:\ kT \leqslant iT_s + \delta < (k + 1)T\} \tag{8}$$

The IDF output for the kth symbol, denoted by A_k, is

$$A_k = \sum_{i \epsilon I^k} y_i \tag{9}$$

The expectation of A_k over the noise, conditioned on $\underline{a}$ and δ, is

$$E[A_k \mid \underline{a}, \delta] = \sum_{i \epsilon I^k} \sum_{k'} a_{k'} q_i(k') \tag{10}$$

To further simplify this expression, define the event indicator function which is 1 if and only if $i \epsilon I^k$, i.e.,

$$l_i(\delta; k) = \begin{cases} 1 & \text{for } kT \leq iT_s + \delta < (k+1)T \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Thus from Eq. (10) we have

$$E[A_k \mid \underline{a}, \delta] = \sum_i \sum_{k'} l_i(\delta; k)\, q_i(k')\, a'_k \tag{12}$$

B. Noise Response

Next we consider the noise response of the IDF to compute the total SNR at the output of the IDF. Let z_i denote the sampled noise response of the filter at time $iT_s + \tau_1$.

$$z_i = \int_{-\infty}^{\infty} n(\zeta) h(iT_s + \tau_1 - \zeta) d\zeta \tag{13}$$

Since the IDF is a linear system, the variance of A_k conditioned on $\underline{a}$ is equal to the variance of the response of the kth symbol due to noise alone, i.e., it is independent of $s(t)$. The variance of A_k is

$$\text{var}[A_k \mid \underline{a}, \delta] = \sum_i \sum_j l_i(\delta; k)\, l_j(\delta; k)\, E[z_i z_j] \tag{14}$$

Note that this variance does depend on δ and k, because the number of samples occurring in the kth symbol depends on δ. Using Eq. (3) and noting that $E[n(t)n(\tau)] = N_0/2\ \delta_0(t - \tau)$, where here $\delta_0(\cdot)$ is the Dirac delta function, we have

$$E[z_i z_j] = R_z(i-j) = \frac{N_0}{2} \int_{-\infty}^{\infty} h((i-j)T_s - \xi) h(\xi) d\xi \tag{15}$$

where R_z is the autocorrelation of z_i.

III. Definition of Signal-to-Noise Ratio Loss

The analog IDF of Fig. 1a is the optimum matched when $p(t) = 1$ for $0 < t < T$ and zero otherwise. The SNR is defined at the IDF output as the ratio of the square of the mean to the variance. Denoting SNR_A for the analog IDF, it is well known [10] that

$$\text{SNR}_A = \frac{2A^2 T}{N_0} \tag{16}$$

We assume with no loss of generality that the signal amplitude $A = 1$, so $\text{SNR}_A = 2T/N_0$.

The SNR at the output of the digital IDF is denoted by SNR_D. We compare the performance of the digital IDF with the analog IDF by considering the ratio

$$\gamma \triangleq \frac{\text{SNR}_D}{\text{SNR}_A} \tag{17}$$

We define SNR_D at the output of the IDF as the ratio of the square of the conditional mean to the conditional variance of the output, conditioned on a given sequence and offset value. From this definition of SNR,

$$\text{SNR}_D = \frac{(E[A_k \mid \underline{a}, \delta])^2}{\text{var}[A_k \mid \underline{a}, \delta]}$$

Then we have

$$\gamma = \frac{N_0}{2T} \text{SNR}_D \tag{18}$$

In the remaining sections, $\gamma_{\text{dB}} = 10 \log_{10}(\gamma)$ (dB) is computed for various filters and data patterns. Normally $\gamma \leq 1$, because the digital IDF has a loss with respect to the analog IDF. The degradation or loss in decibels is the negative of γ_{dB}, and minimum loss corresponds to the maximum attainable γ, which typically approaches one ($\gamma_{\text{dB}} = 0$ dB). Maximum loss is unbounded. In some cases, for a given data pattern $\underline{a}$, there

is a gain in SNR (e.g., all-ones sequence), in which cases $\gamma > 1$ and the degradation is negative in decibels.

IV. SNR Performance for Special Cases

In general, the pulse shape $p(t)$ may be chosen to take numerous shapes (e.g., raised root-cosine). In some cases, it is chosen to extend over more than one symbol duration, e.g., for partial response signaling (sometimes referred as correlated coding or controlled intersymbol interference). For bandwidth-limited channels, the pulse shape and duration are selected to increase the bandwidth efficiency of the communication system.

We consider only non-overlapping rectangular pulses throughout the rest of this article, since this pulse shape has traditionally been used for NASA's deep space missions. It is pointed out that the results of the previous sections, namely the expressions for average signal response in Eq. (12) and the noise variance in Eq. (15), hold regardless of the pulse shape or its duration.

In the case of the rectangular pulse we simply have

$$p(t) = \begin{cases} 1 & 0 < t \leqslant T \\ 0 & \text{otherwise} \end{cases}$$

then from Eq. (6)

$$q_i(k) = \int_0^T h(iT_s - kT + \delta - x)dx \qquad (19)$$

and

$$E[A_k | \underline{a}, \delta] = \sum_i \sum_{k'} l_i(\delta; k)\, q_i(k')a_{k'}$$

It is useful to write Eq. (19) as

$$q_i(k) = \int_{kT}^{(k+1)T} h(iT_s + \delta - x)dx \qquad (20)$$

We now consider two different low-pass filters, one causal and one non-causal.

A. First Order Low-Pass Filter (Causal)

The impulse response for a first order low-pass filter is $h(t) = We^{-Wt}u(t)$, where $u(t)$ denotes the unit step response and W is the radian cutoff frequency of the filter. The filter in this case is causal and therefore physically realizable, i.e., $h(t) = 0$ for $t < 0$. Using Eq. (20) to evaluate $q_i(k)$ from Eq. (12),

$$q_i(k) = \begin{cases} (1 - e^{-W(iT_s + \delta - kT)}) & ;\ kT < iT_s + \delta < (k+1)T \\ e^{-W(iT_s + \delta - T)} & ;\ (k+1)T < iT_s + \delta \\ 0 & ;\ iT_s + \delta < kT \end{cases} \qquad (21)$$

To derive the equation for the noise variance, we use the expression in Eq. (15), resulting in

$$\text{var}[A_k | \delta] = \frac{WN_0}{2} \sum_{i \in I^k} \sum_{j \in I^k} e^{-|i-j|T_s W} \qquad (22)$$

To obtain the relative performance of the digital IDF when using a first order low-pass filter, we use the signal response expression in Eq. (21) and the noise variance in Eq. (22). Then, from Eq. (18), γ is

$$\gamma = \frac{\left(\sum_{i \in I^k} \sum_k a_k q_i(k)\right)^2}{WT\left(\sum_{i \in I^k} \sum_{j \in I^k} e^{-|i-j|T_s W}\right)} \qquad (23)$$

where $q_i(kT)$ is defined in Eq. (21). For a known data pattern, Eq. (21) may be evaluated for different k, δ, and time bandwidth products (WT).

B. Ideal Low Pass Filter (Non-Causal)

The ideal low-pass filter with unit gain and low-pass bandwidth W Hz is non-causal with the impulse response

$$h(t) = 2W \frac{\sin 2\pi W t}{2\pi W t} = 2W \sin c(2\pi W t) \qquad (24)$$

The expression for the signal response in Eq. (20) becomes

$$q_i(k) = \frac{1}{\pi} \int_{kT}^{(k+1)T} \frac{\sin 2\pi W(iT_s + \delta - x)}{(iT_s + \delta - x)} dx \qquad (25)$$

It is possible to express Eq. (25) in terms of

$$Si(x) = \int_0^x \frac{\sin u}{u}\, du$$

as

$$q_i(k) = \frac{1}{\pi}\left[Si(2\pi W(iT_s - (k+1)T + \delta)) - Si(2\pi W(iT_s - kT + \delta))\right] \quad (26)$$

To find the noise variance, it suffices to note that the noise spectral density at the output of the filter is

$$S_n(f) = \begin{cases} \frac{N_0}{2} & \text{for all } |f| < W \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

and thus the autocorrelation function is

$$R_n(\tau) = N_0 W \frac{\sin 2\pi W\tau}{2\pi W\tau} \quad (28)$$

Thus, the noise variance at the output of the IDF can be expressed from Eq. (28) as

$$\text{var}[A_k \mid \delta] = \frac{N_0}{2} \sum_{i \in I^k} \sum_{j \in I^k} \sin c(2W(i-j)T_s) \quad (29)$$

The performance γ can be evaluated from Eq. (18), which yields

$$\gamma = \frac{\left[\sum_{i \in I^k} \sum_{k'} a_{k'} \frac{1}{\pi}\left(Si(2W\pi(iT_s - (k'+1)T + \delta)) - Si(2W\pi(iT_s - k'T + \delta))\right)\right]^2}{WT \sum_{i \in I^k} \sum_{j \in I^k} \sin c(2W(i-j)T_s)} \quad (30)$$

V. Numerical Results

Numerical results are presented only for the ideal low-pass filter. The motivation to consider the ideal low-pass filter is to eliminate aliasing in an ideal manner. The use of a realizable filter such as that of Butterworth or Chebyshev [11] and [12] does not greatly influence the results, since the realizable filter can be considered an approximation to the unrealizable filter with finite group delay [12].

All the computational results were obtained by explicitly evaluating Eq. (30) and Eq. (23) for arbitrary input signal sequences a. The $Si(\cdot)$ function in Eq. (30) was implemented using a 500 point look-up table provided in [9].

The simulation result for the first order low-pass filter SNR [Eq. (23)] is not included in this article. It was, however, computed by the authors to verify the results against Natali's results [1], as was pointed out in the introduction.

For Figs. 4 through 13, the letters a through f correspond to the following time bandwidth products (WT): $a = 2$, $b = 1.75$, $c = 1.50$, $d = 1.0$, $e = 0.75$, and $f = 0.5$, unless stated otherwise.

A. Spectral Analysis

To gain a better understanding of the digital IDF, we investigated the spectral properties of the signals processed in the IDF. We considered two cases. For the first case, the data signal $s(t)$ is the alternating +1, −1 sequence with the Fourier series

$$s(t) = \frac{4}{\pi} \sum_{i=0}^{\infty} \frac{(-1)^{i+1}}{2i+1} \cos 2\pi f_0(2i+1)t \quad (31)$$

where $f_0 = 1/2T$. We refer to this case as case 1. Note that the signal has energy only at odd harmonics of f_0.

For case 1 and the ideal filter, the bandwidth W determines the number of odd harmonics that pass through the filter. When the time bandwidth product WT is greater than 1.5 and less than 2.5, the first two odd harmonics pass through the filter; and when $0.5 \leqslant WT < 1.5$, only the first harmonic is passed through the filter.

The second case, case 2, is the binary random waveform with rectangular pulse shape, which has the autocorrelation function [13] $R_s(\tau)$:

$$R_s(\tau) = \begin{cases} 1 - \frac{|\tau|}{T} & |\tau| < T \\ 0 & \text{otherwise} \end{cases} \quad (32)$$

with power spectral density

$$S_s(f) = T\frac{\sin^2(\pi fT)}{(\pi fT)^2} \tag{33}$$

The noise process $n(t)$ and the transmitted signal $s(t)$ are mutually independent. The autocorrelation function of the observed signal $r(t)$ is

$$R_r(\tau) = R_s(\tau) + R_n(\tau) \tag{34}$$

The spectral shapes of the transmitted signal $s(t)$ and the noise process $n(t)$ are shown for case 2 in Fig. 3. The time-bandwidth product WT determines the number of lobes of the signal spectrum Eq. (33) passed through the filter.

B. Digital IDF Output Noise Variance

The noise variance at the output of the IDF for the case when there are N samples in the symbol is

$$\text{var}[A_k \mid \underline{a}, S] = NR_n(0) + \sum_{i=1}^{N-1} 2(N-i)R_n(iT_s) \tag{35}$$

This variance reduces the $NR_n(0)$ when the samples are independent. This is possible by choosing the sampling period $T_s = 1/2W$, which is evident from Eq. (28).

C. Performance Loss Versus Offset and Bandwidth for Known Waveforms

In this section, we consider known signals as input to the digital IDF. The output of the ideal low-pass filter depends on both past and future inputs. To approximate this, we consider 21-symbol blocks, and the 11th symbol is analyzed for each data pattern. A block of 21 symbols was found to be sufficiently long to analyze the IDF for different data patterns. This is reaffirmed in the following section (V.C.1) by considering the spectrum of the sampled waveform for the alternating data pattern. Five different data patterns are considered:

(1) Alternating data pattern: $\underline{a} = (-1, +1, -1, +1, -1, \ldots)$

(2) Single pulse: $\underline{a} = (-1, -1, -1, \ldots, -1, +1, -1, \ldots)$

(3) Two ones: $\underline{a} = (-1, +1, -1, +1, \ldots, +1, -1, +1, +1, -1, +1, \ldots)$

(4) Three ones: $\underline{a} = (+1, -1, +1, -1, \ldots, +1, -1, +1, +1, +1, -1, +1, -1, \ldots)$

(5) All ones: $\underline{a} = (+1, +1, +1, \ldots)$.

We analyze the waveform and SNR for each pattern for different WT, particularly for when there are four samples per symbol, i.e., $T_s = T/4$.

In the following sections and related figures, the offset δ is defined as the length of time from the start of the 11th symbol to the time when the first sample of the 11th symbol occurs.

1. Pattern 1, alternating data. The sampled waveform for the 11th symbol, a −1 pulse, is shown in Fig. 4(a) for $WT = 2$. In Fig. 4(a), for every sampling offset value, with increments 0.05 T_s, a unique English letter (a through t) is used to indicate the point at which the sample occurs. Every letter occurs four times, corresponding to the four samples per symbol.

In Fig. 4(b), the anti-aliasing filter output waveform is shown for different time bandwidth products. The sampled waveform approaches a sinusoid when $0.5 \leqslant WT \leqslant 1.5$, and approaches the sum of a sinusoid and its third harmonic when $1.5 \leqslant WT \leqslant 2.5$. This indicates that the sequence of 21 data symbols is sufficiently long to approximate the infinite alternating sequence, since the waveforms agree with harmonic properties discussed earlier in Section V.1.

The degradation is depicted in Fig. 4(c) for different offset values, when WT ranges from 0.75 to 2.0. For all WT cases, the worst case occurs when $\delta = 0$, with the value of the worst case ranging from 1.6 dB at $WT = 0.75$ to 1.96 dB at $WT = 1.75$. To find out how much loss is due to sampling, the loss for $T/T_s = 20$ is shown in Fig. 4(d). As expected, since the sampling rate is high, the loss depends mainly on WT, not on δ. For the case when $WT = 1$, the worst case loss is decreased from 1.92 dB at $T_s = T/4$ to 1.10 dB for $T_s = T/20$. When $T_s = T/20$, the digital IDF almost approximates the analog IDF (with finite bandwidth). Hence, it can be deduced that about 0.82 dB loss results from sampling with $T_s = T/4$, and 1.1 dB loss results from the bandwidth limiting of the received signal.

2. Pattern 2, single pulse. We consider here the second pattern and compare it to the first pattern. The results are compared to those of Hartman [2], who considered the same pattern.

Hartman points out that when the time-bandwidth product is an even integer, the maximum loss results when the binary waveform is a single 1 preceded and followed by all −1's (or vice versa). He also shows that when the time-bandwidth product is an odd integer the maximum loss results when the binary waveform consists of alternating 1's and −1's, respectively.

In Fig. 5(a) the resulting waveform for this sequence is shown for different WT values. In Fig. 5(b) the degradation is

shown as a function of the offset δ. Performances of the alternating data pattern and a single pulse are compared in Fig. 6, for $WT = 2$. We find that the loss is virtually the same for the two sequences, and the alternating sequence results in slightly more degradation, approximately 0.01 dB for the same values of offset. This is in slight disagreement with Hartman [2]. For $WT = 1$, the worst case loss is 1.77 dB for the single pulse, and 1.92 dB for the alternating pattern, which agrees with Hartman [2].

3. **Pattern 3, an asymmetric pattern.** The third data pattern is (-1, +1, . . ., +1, -1, +1, +1, -1, +1, . . .) considered. The filtered output waveforms shown in Fig. 7(a) and Fig. 7(b) depict the corresponding degradation. Since the waveform is asymmetrical about the $T/2$ point, the loss is also asymmetrical about the $T_s/2$ point.

4. **Pattern 4, alternating except for three ones.** The loss is shown for this pattern in Fig. 8. The sampled waveform is not shown in this case, since it is almost constant throughout the symbol time. The performance curve in this case is symmetrical, since the symbol waveform is symmetrical around the $T/2$ point. In this case, the worst case loss does not occur for $\delta = 0$ or $\delta = T_s$, which is different from the previous cases. Also, for cases b, c, and d ($WT = 1.75$, 1.5, and 1.0), there is a gain rather than a loss in SNR for all or most offsets. This occurs because the ISI happens to aid in these cases.

5. **Pattern 5, all ones.** Cases 3 and 4 result in a slight gain in the SNR for certain offsets and time bandwidth products. To investigate this point, the loss curve in Fig. 9 is shown, when a step function (all 1's pattern) is used as the symbol sequence. In this case when $WT < 1$ is selected, the performance curve indicates a constant gain. This is a consequence of the expression for γ in Eq. (30). By decreasing the anti-aliasing filter bandwidth to $WT < 1$, the amount of filtered noise power decreases. The signal power does not significantly decrease for long periods of constant data. This results in a performance gain for these cases.

D. Selection of Bandwidth and Sampling Rate

In implementation of the digital IDF, for a given symbol rate T, a reasonable criterion is to select the sampling period T_s and the bandwidth W such that the average loss is minimized, where the average is over the ensemble of all possible data patterns. One could consider either averaging the loss over all offsets or the worst case offset. Both bandwidth and sampling may also be restricted by hardware or other considerations.

It is well known, when applying the orthonormal set of radial prolate spheroidal functions, to expand a band-limited signal in the function space of finite energy signals, that the number of orthogonal dimensions (eigenvalues) necessary to describe a band-limited function over a T-sec time interval is approximately $2WT$ [12].

The Nyquist sampling theorem [11] requires that the sampling period satisfy the inequality

$$T_s \leqslant \frac{1}{2W} \tag{36}$$

or, equivalently, the number of samples/symbol must satisfy the inequality

$$\frac{T}{T_s} \geqslant 2WT \tag{37}$$

The digital IDF asymptotically approaches the analog IDF when W, and T/T_s, approach infinity. This leads to an infinite time-bandwidth product which is unrealizable in a physical system.

The bandwidth W should be selected such that most of the signal power is passed through the filter. Furthermore, it is clear from Eq. (36) that as the bandwidth is increased, the required sampling period decreases, and more samples per symbol are required.

To select the best filter bandwidth given a symbol rate, one must also consider the current hardware constraints of technology. The two most important constraints are the speed and accuracy of available A/D converters, and the speed and accuracy of signal processing hardware such as multipliers and accumulators. Therefore, the sampling rate and the filter bandwidth must be selected such that the hardware is practically implementable.

1. **Optimum choice for T_s and W.** Given a symbol rate, the necessary steps for the design engineer are:

(1) To determine the maximum practical limit to the sampling rate or, equivalently, the maximum number of samples that is possible during each symbol time.

(2) Once the sampling period T_s is determined, then the bandwidth W must be selected such that the inequality [Eq. (37)] is satisfied, preferably with equality. This results in passing maximum signal power into the digital IDF, while satisfying the sampling criterion.

Investigating the SNR loss when Eq. (37) is satisfied with strict equality is the subject of the study in the following section.

2. Monte Carlo simulation. We consider the class of symbol patterns that are random binary sequences selected from an equally probable binary alphabet. In Figs. 10 through 13, the average loss for binary vectors of length 4620 bits, generated from a binary symmetric source, is computed for various time-bandwidth products and sampling rates.

The averages are computed by breaking the 4620 bit vector into 220 blocks of 21 symbols each and computing the loss of the 11th symbol for each block. It was confirmed previously that 22 symbols are sufficiently long to approximate an infinite length sequence, for simulation purposes, when the 11th symbol is being analyzed.

The stopping rule for computing the average loss is to repeat computing the loss for every block until the accumulated loss does not change by more than 10^{-4} for 10 consecutive blocks. At this point the average loss is computed by dividing the accumulated loss by the number of blocks that have been processed. This approach has the advantage of averaging over the ensemble of ISI patterns. Furthermore, the incremental loss computed for a given block of data patterns is independent of the loss computed for the previous data pattern.

The results confirm that the minimum loss averaged over the ensemble of all possible test patterns is achieved when δ, the sampling offset, is $T_s/2$. This conclusion is only true if the sampling rate is an integer multiple of the symbol rate. This is clarified when we consider the case in which the symbol period is not an integer multiple of the sampling period.

Figure 10(a) depicts the average loss versus offset for $T_s = T/2$ and for several different WT products. For $WT = 1$, the worst case SNR loss in this case is about 2.4 dB, at $\delta = 0$. When $WT > 1$ the inequality of Eq. (37) is violated, because the sampling rate is too low for the bandwidth, and aliasing occurs. The curves for $WT > 1$ are depicted to exhibit the loss due to aliasing. Only curves *d* and *e*, with $WT = 1$ and 0.75, satisfy Eq. (37). For the best offset, $\delta = 0.5$, which corresponds to phase locked sampling, the loss is only 0.7 dB.

Figure 10(b) depicts the case when $T/T_s = 4$. In this case Eq. (37) is satisfied for all WT considered. For most values of the offset, the loss decreases uniformly as the time bandwidth product approaches 2. Thus, selecting $W = 2/T$ is the most appropriate choice, which satisfies Eq. (37) with strict equality. The worst case loss occurs when $\delta = 0$, and the corresponding loss is approximately 1.2 dB for all WT cases. The widest bandwidth is recommended because it minimizes the average loss, averaged over all offsets.

Figure 10(c) depicts the case when $T/T_s = 8$. In this case the minimum loss (at $\delta = T_s/2$) is virtually unchanged from the case when $T/T_s = 4$. This loss is approximately 0.3 dB to 0.4 dB for $WT = 1.5$ and 2.0. For the widest bandwidths, $WT = 1.5$ and 2.0, the worst case loss is improved by 0.6 dB, from 1.2 dB to 0.6 dB, over the case when $T_s = T/4$. This improvement results from doubling the number of samples per symbol at $WT = 2$.

The conclusion is that sampling at twice the Nyquist rate reduces the worst case loss by 0.6 dB compared to sampling at the Nyquist rate, for $WT = 2$. Sampling 16 times the symbol rate with $WT = 4$ [see Fig. 11(a)] results in the worst case loss of 0.3 dB, of which 0.2 dB is due to bandwidth limiting and only 0.1 dB is due to sampling and the IDF.

Two cases are considered to investigate the case when WT is selected to be greater than two. In Fig. 11(a) with $WT = 4$, and in Fig. 11(b) with $WT = 8$, the average degradation is shown for different T/T_s ratios. As the number of samples/symbol is increased in Fig. 11(a), the variation in the loss due to the offset becomes negligible and both the worst case and best case losses approach approximately 0.2 dB. This indicates that the degradation due to the offset becomes negligible as the number of samples per symbol is increased.

Regressing to the practical scenario, suppose that given practical hardware constraints, it is possible to obtain a maximum of eight samples per symbol. To find the best choice for the bandwidth, we have compared the loss for $WT = 2$ and 4 in Fig. 12. It is apparent that in the worst case offset ($\delta = 0$), both cases result in equal loss of 0.6 dB, but the minimum degradation is better by 0.2 dB when $WT = 4$. Thus, this choice of bandwidth is better when $T_s = T/8$.

3. Summary of Monte Carlo simulation results. To summarize the results for simulation, it is concluded that $0.75 < WT < 1$ is appropriate when $T_s = T/2$. When $T_s = T/4$ it is recommended to use the widest filter that eliminates aliasing, i.e., $W = 2/T$. Then sampling is at exactly the Nyquist rate. This is better than selecting $WT < 2$ and sampling higher than the Nyquist rate. When eight samples per symbol are available, i.e., $T_s = T/8$, it was deduced that the worst case loss is the same for $W = 2/T$ and $W = 4/T$. However, the minimum average loss is lower for wider bandwidth, $W = 4/T$. When the sampling rate is limited, the widest bandwidth that satisfies Eq. (37) should be used. Sampling at twice the Nyquist rate is recommended when the bandwidth is limited.

E. Effects of Asynchronous Sampling

Normally, in practical systems the sampling clock is not an exact multiple of the symbol clock. This occurs when the sampling is at a fixed rate and the symbol rate varies, for example, with Doppler shift.

Figure 13 depicts the average loss for a random data pattern when $T/T_s = 4.5$. Some symbols have four samples, and some have five. When the number of samples is four (or when the offset $\delta < T_s/2$), the degradation is approximately 0.3 dB less than when the number of samples is five. This is true for all WT considered. The average loss for five samples per symbol is due to additional noise power in the last sample, since the last sample occurs near the transition point of the filtered symbol waveform. Despite this variation in loss with the number of samples, the worst case loss for $T/T_s = 4.5$ is only 0.7 dB for $WT = 1.5$ and 2.0, which is less than the worst case loss of 1.2 dB when $T/T_s = 4$. Even though there is a variation in loss with number of samples per symbol, a higher sampling helps performance rather than hurting it.

VI. Summary and Conclusions

The performance of the digital IDF was studied in this article, with special attention to the effect of offset in sampling. We derived the expressions for the signal response, noise variance, and IDF output SNR for general pulse shapes and anti-aliasing filters. These results were then specialized to the rectangular pulse shape and the ideal low-pass filter. Performance was evaluated in terms of loss in SNR compared to an analog IDF with infinite bandwidth.

Performance was evaluated for five different known data sequences, and for random sequences. Observations regarding the worst case degradation and the effects of the offset in the sample time were made. It was concluded that when the system is bandlimited, i.e., the time bandwidth product WT is finite, a sampling rate in excess of the Nyquist rate ($2W$) should be chosen. For example, when the bandwidth is only twice the symbol rate, sampling at twice the Nyquist rate results in a 0.6 dB improvement over sampling at the Nyquist rate, when the worst case loss is the performance criterion. When the sampling rate is limited, i.e., T/T_s is fixed the bandwidth W should be selected such that $W = 1/2T_s$.

The above results are for the case when the sampling rate is not phase locked to the symbol rate. For an integer number of samples per symbol, it was shown that the degradation due to offset in sampling is minimized when the offset is half the sampling period, i.e., $\delta = T_s/2$. For four samples per symbol, the worst case loss is 0.9 dB greater than the best case. This means that phase locked sampling is 0.9 dB better than the worst case for the non-phase-locked sampling, when the ratio of symbol rate to the sampling rate is small ($\leqslant 4$).

The loss due to the offset becomes negligible when the number of samples/symbol becomes sufficiently large. With $WT > 4$, $T_s < 1/2W$, the digital IDF always performs within 0.6 dB of the analog IDF, for the random data pattern and the worst case sampling offset.

The effect of a non-integer ratio of sampling rate to symbol rate was also studied, for the case when $T/T_s = 4.5$. For a given time bandwidth product, the worst case loss is lower than the case when $T/T_s = 4$. Thus letting the sampling rate be a non-integer rate relative to symbol rate did not degrade the performance.

VII. Direction for Future Research

For the Deep Space Network receiver, the losses incurred for the digital IDF at high symbol rates are undesirable. To reduce the loss, it is possible to weight each sample in the IDF such that the loss is minimized. That is, instead of performing a simple summation, a weighted summation is performed. This will be considered in a later article.

Glossary of Terms

T_s – Sampling time in seconds

T – Symbol time in seconds

W – Filter bandwidth in hertz

a_i – Transmitted symbol

$p(t)$ – Pulse shaping waveform

A_k – The sampled output of the Integrate-and-Dump Filter at time k

y_i – The sampled output of the pre-filter

$s(t)$ – The transmitted signal

$n(t)$ – Additive white Gaussian noise with flat spectral density $N_0/2$

$r(t)$ – The received signal

$y(t)$ – The output of the low-pass prefilter

$h(t)$ – The transfer function of the low-pass prefilter

τ_0 – Transport lag from transmitter to receiver in seconds

τ_1 – Sampling delay in seconds

$R_x(\tau)$ – Autocorrelation function of signal $x(t)$

References

[1] F. D. Natali, "Comparison of Analog and Digital Integrate-and-Dump Filters," *Proc. IEEE*, vol. 57, pp. 1766–1768, October 1969.

[2] H. P. Hartman, "Degradation of Signal-to-Noise-Ratio due to IF Filtering," *IEEE Trans. Aero. and Elect. Sys.*, vol. AES-5, no. 1, pp. 22–32, January 1969.

[3] G. L. Turin, "An Introduction to Digital Matched Filters," *Proc. IEEE*, vol. 64, pp. 1092–1112, July 1976.

[4] T. L. Lim, "Non-Coherent Digital Matched Filters: Multibit Quantization," *IEEE Trans. Comm.*, vol. COM-26, pp. 409–419, April 1978.

[5] C. M. Chie, "Performance Analysis of Digital Integrate-and-Dump Filters," *IEEE Trans. Comm.*, vol. COM-30, no. 8, pp. 1979–1983, August 1982.

[6] C. W. Helstrom, "Calculating Error Probabilities for Intersymbol and Cochannel Interference," *IEEE Trans. Comm.*, vol. COM-34, no. 5, pp. 430–435, May 1986.

[7] A. J. Levy, "Fast Error Rate Evaluation in the Presence of Intersymbol Interference," *IEEE Trans. Comm.*, vol. COM-33, no. 5, pp. 479–481, May 1985.

[8] C. M. Chie, "The Effects of Transmitter/Receiver Clock Time-Base Instability on Coherent Communication System Performance," *IEEE Trans. Comm.*, vol. COM-30, no. 3, pp. 510–516, March 1982.

[9] M. Abramowitz and A. I. Stegun (eds.), *Handbook of Mathematical Functions*, Washington, D.C.: National Bureau of Standards; 1964.

[10] D. A. Whalen, *Detection of Signals in Noise*, New York: Academic Press, 1971.

[11] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, New York: Prentice-Hall, 1975.

[12] A. Papoulis, *Signal Analysis*, New York: McGraw-Hill, 1977.

[13] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, New York: McGraw-Hill, 1965.

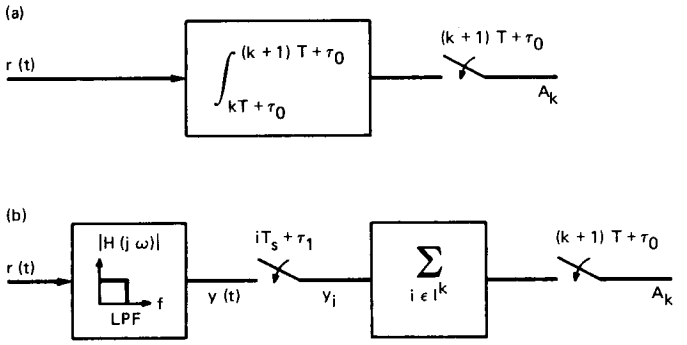


Fig. 1. Analog and digital Integrate and Dump Filter (IDF): (a) analog IDF; and (b) digital IDF

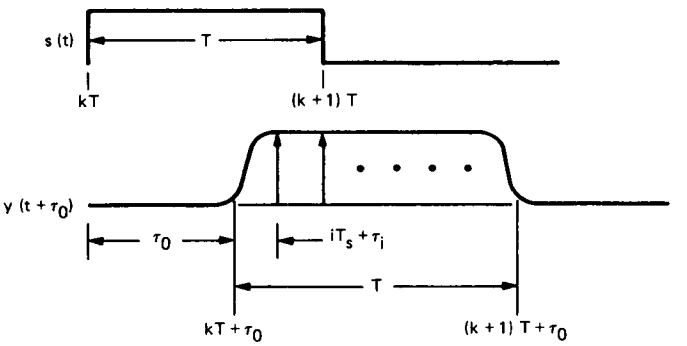


Fig. 2. Offset in sampling

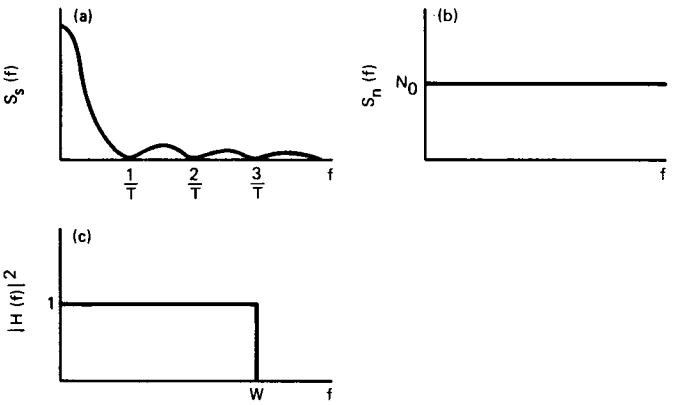


Fig. 3. Signal and noise spectral density for binary random waveform: (a) spectral density of transmitted signal; (b) spectral density of the noise; and (c) frequency response, the ideal filter

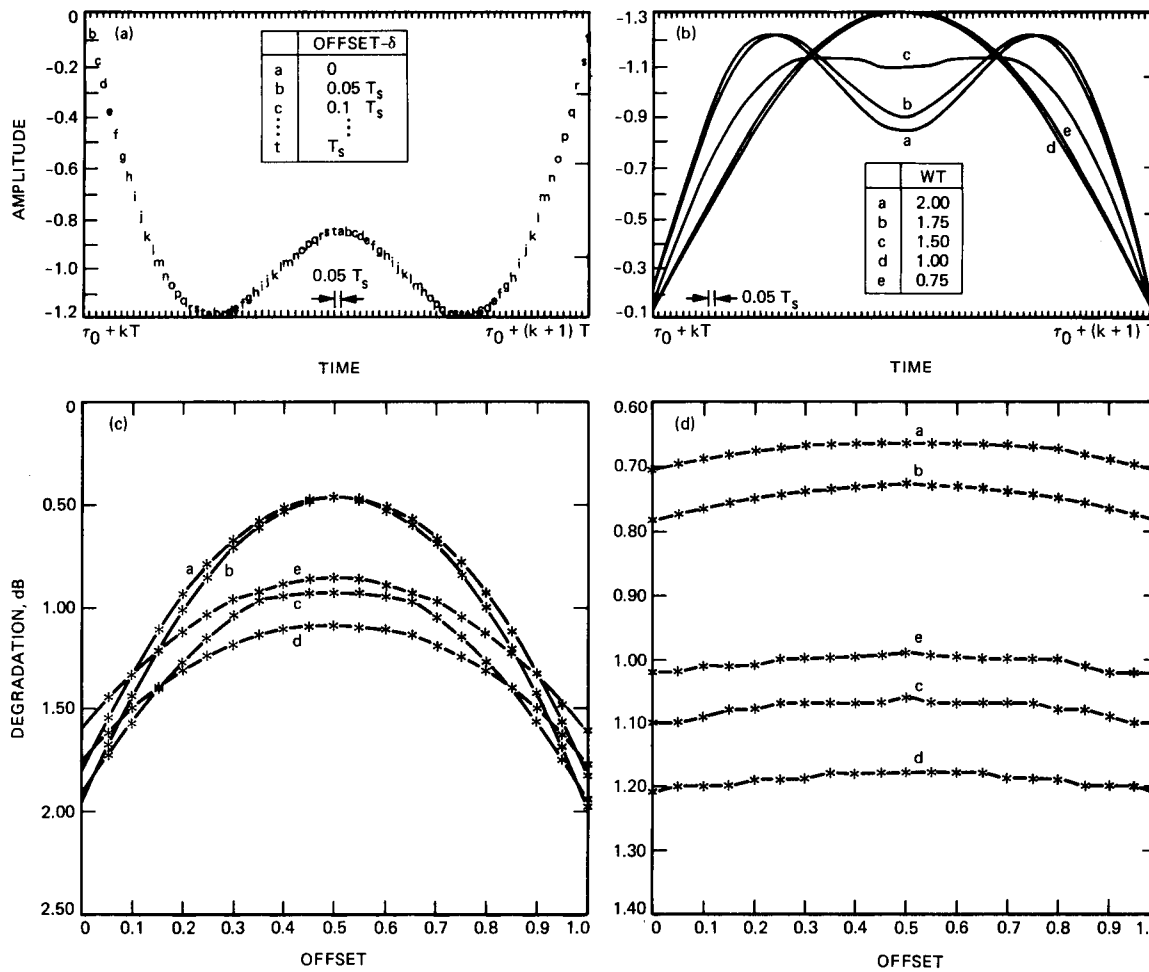


Fig. 4. Results for alternating data pattern: (a) sampled waveform; (b) filter output for different *WT*; (c) alternating data pattern $T/T_S = 4$; and (d) alternating data pattern $T/T_S = 20$

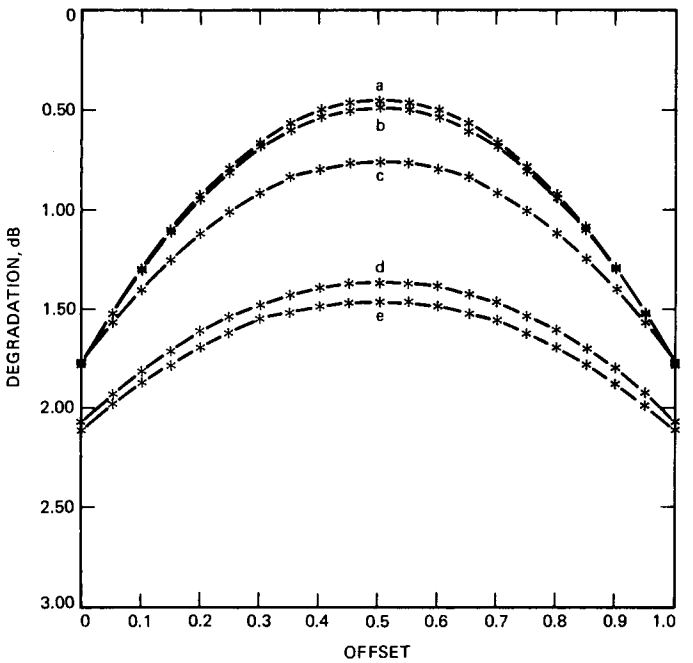


Fig. 5. Degradation vs. offset: single pulse data pattern $T/T_s = 4$

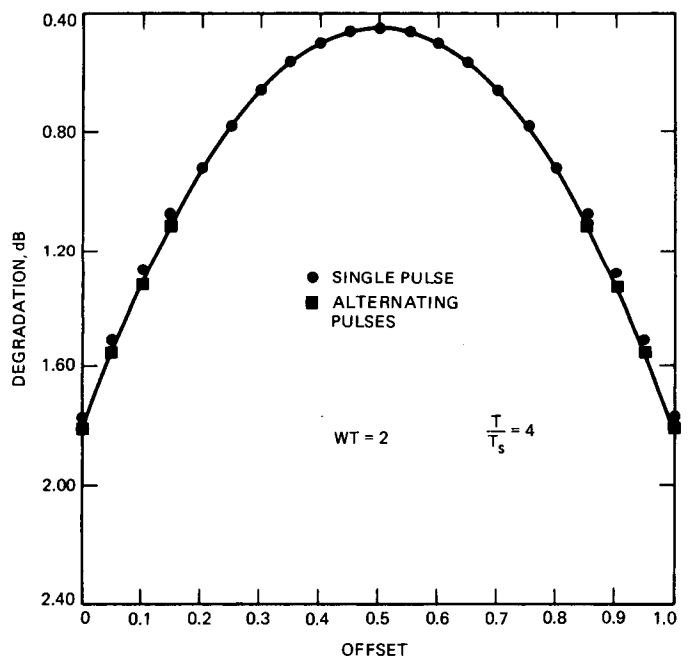


Fig. 6. Degradation comparison vs. offset for alternating and pulse data pattern

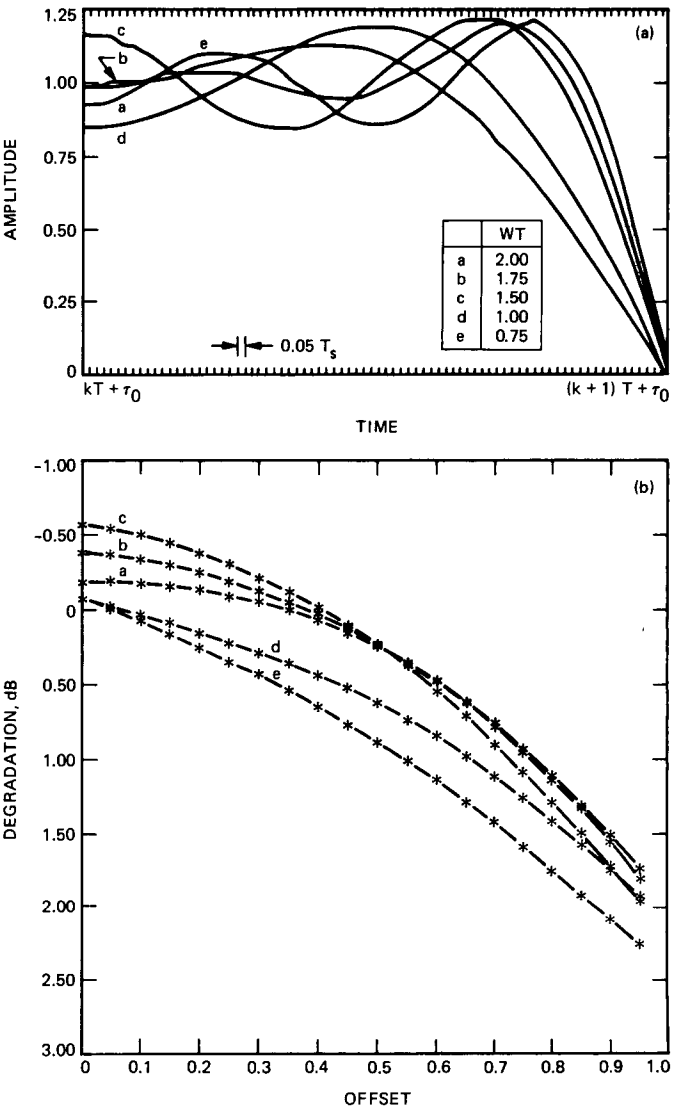


Fig. 7. (a) Signal output for different *WT* for data pattern -1, +1, -1, +1, . . . +1, -1, +1, +1, -1, +1 . . . ; (b) Degradation vs. offset two-ones data pattern

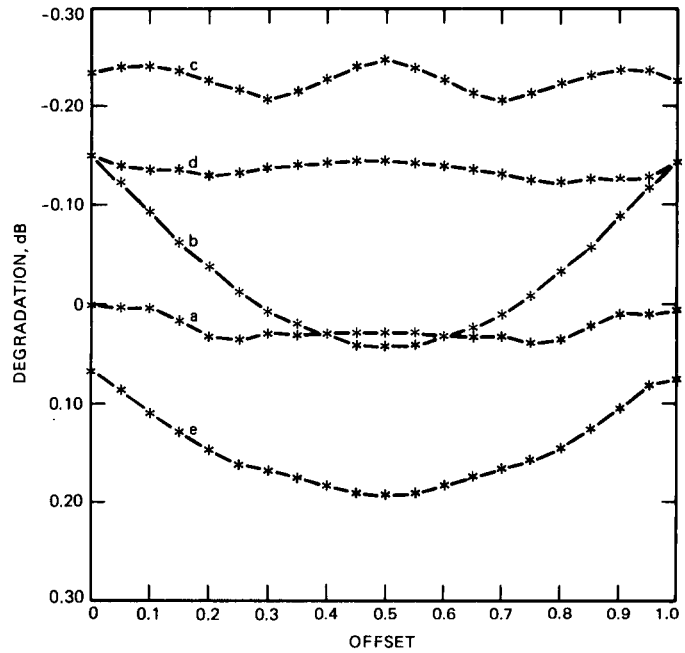


Fig. 8. Degradation vs. offset three-ones data pattern

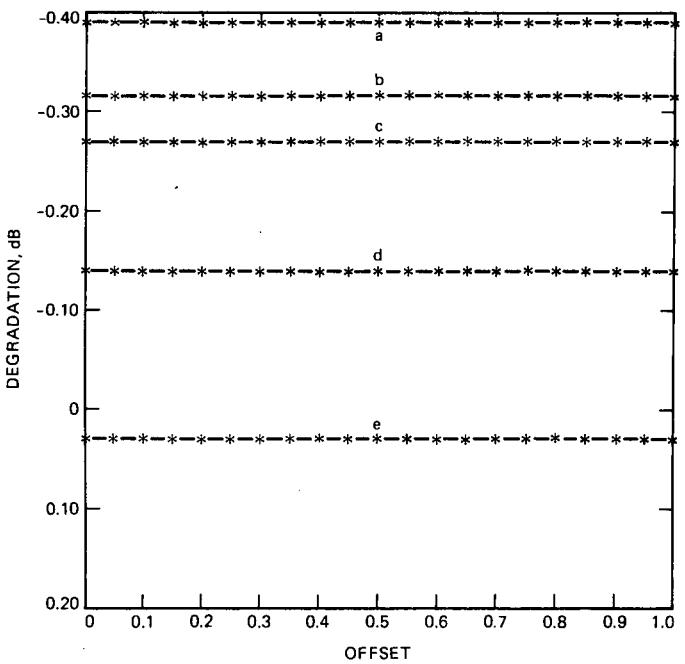


Fig. 9. Degradation vs. offset all ones data pattern

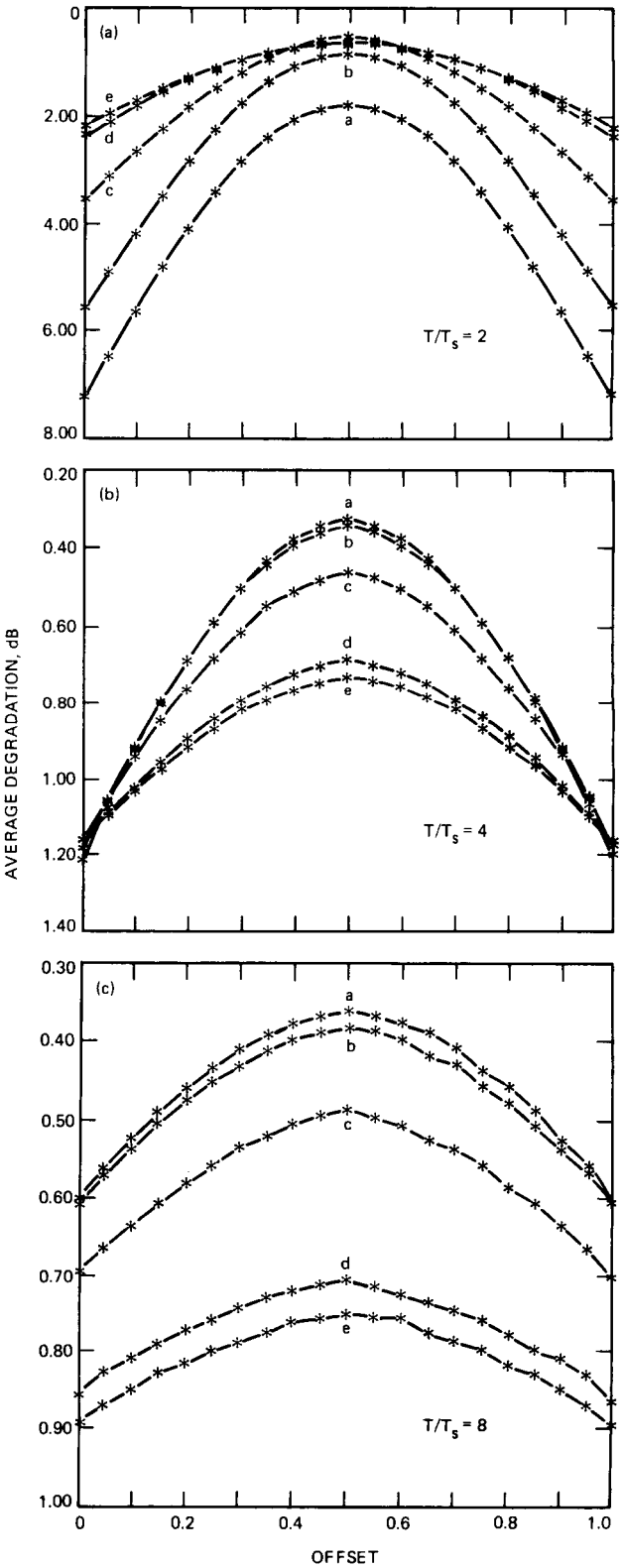


Fig. 10. Average degradation vs. offset: (a) $T/T_S = 2$; (b) $T/T_S = 4$; and (c) $T/T_S = 8$

C-3

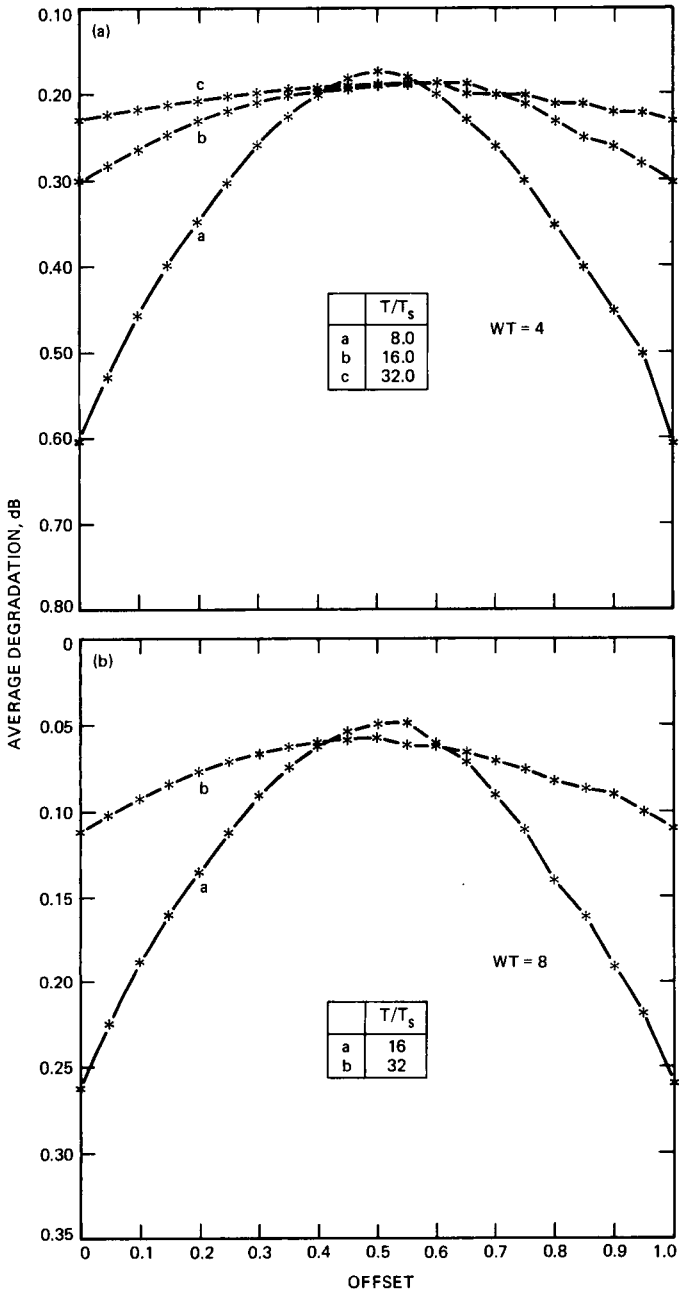


Fig. 11. Average degradation vs. offset: (a) *WT* = 4 and (b) *WT* = 8

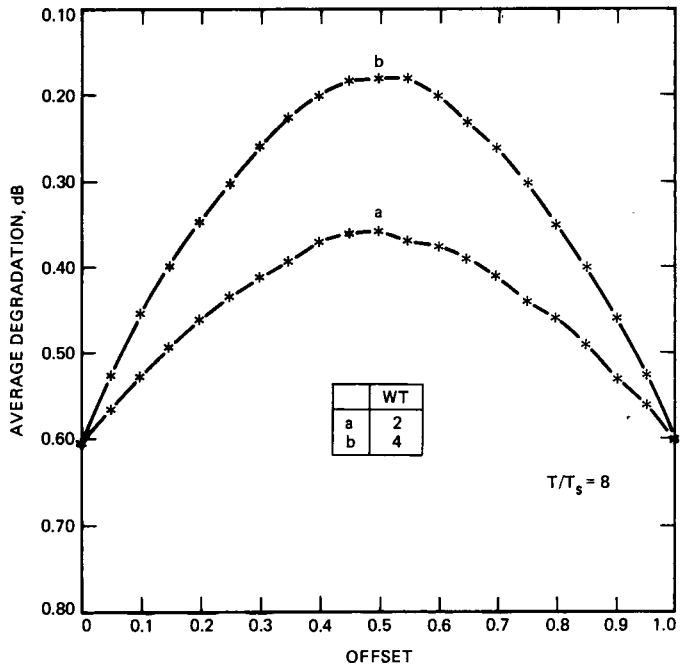


Fig. 12. Average degradation vs. offset (T/T_S = 8)

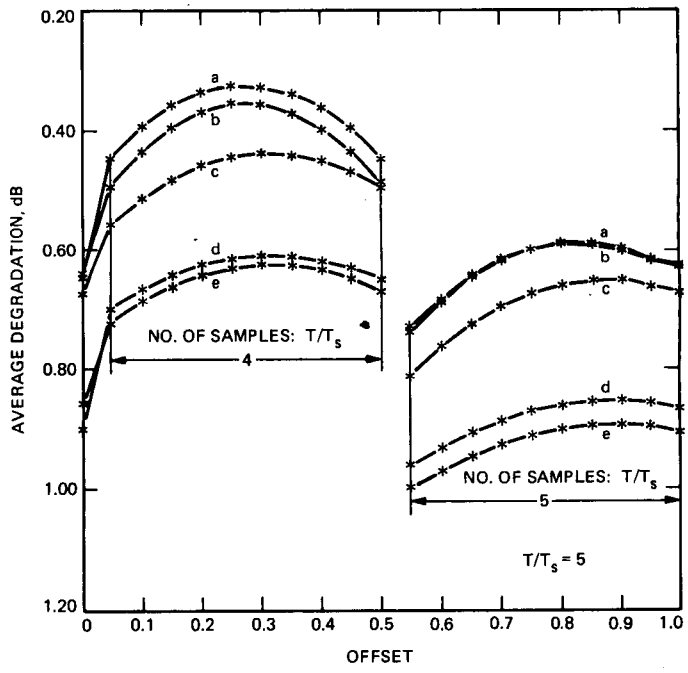


Fig. 13. Average degradation vs. offset (T/T_S = 4.5)

Detection of Signals by Weighted Integrate-and-Dump Filter

R. Sadr
Communications Systems Research Section

A Weighted Integrate-and-Dump Filter (WIDF) is presented that results in reducing those losses in telemetry symbol SNR which occur in digital Integrate-and-Dump Filters (IDFs) when the samples are not phase locked to the input data symbol clock. The Minimum Mean Square Error (MMSE) criterion is used to derive a set of weights for approximating the analog integrate-and-dump filter, which is the matched filter for detection of signals in additive white Gaussian noise. This new digital matched filter results in considerable performance improvement compared to unweighted digital matched filters. An example is presented for a sampling rate of four times the symbol rate. As the sampling offset (or phase) varies with respect to the data symbol boundaries, the output SNR varies 1 dB for an unweighted IDF, but only 0.3 dB for the optimum WIDF, averaged over random data patterns. This improvement in performance relative to unweighted IDF means that significantly lower sampling and processing rates can be used for given telemetry symbol rates, resulting in reduced system cost.

I. Introduction

The effect of "offset sampling" for the unweighted digital Integrate-and-Dump Filter (IDF) was considered in [1]. A set of practical guidelines is outlined in [1] that can be used to determine the appropriate sampling period and the filter bandwidth for the digital IDF. In addition, the effect of offset sampling was comprehensively studied, and the degradation due to approximating the analog IDF with digital IDF was analyzed. By "offset sampling," we mean that the sampling clock is not phase locked to the telemetry symbol clock.

The IDF is the optimum matched filter for detection of signals in Additive White Gaussian Noise (AWGN). In this article, a new class of digital matched filters is considered which decreases the degradation due to approximating the analog IDF with the digital IDF.

The problem is formulated in the context of waveform tracking. The waveform which is tracked by the linear estimator is the sampled output of the analog IDF. The mean square error criterion is used to derive the digital matched filter. The observed signal for derivation of the digital matched filter is the sampled sequence of the received signal during a single symbol time of T seconds. In a sampled data system, normally an anti-aliasing low-pass (or bandpass) filter is used to filter the analog source. The effect of this filter is specifically considered.

In Section II, the underlying system is described. In Section III, the new digital matched filter is formulated. In Section IV, a linear system is proposed that generates the necessary autocorrelation functions for computation of the optimal weight sequence. In Section V, the average signal response expression

is derived for the weighted integrate-and-dump filter in the presence of offset sampling. In Section VI, the noise response of the system is considered. In Section VII, the definition of SNR loss due to the approximation of the analog IDF with the digital IDF is stated. In Section VIII, results of the previous sections are used to find the optimum weighted IDF for a special case when an ideal filter is used prior to sampling the observed signal and the transmitted signal is a sequence of rectangular pulses. The relationship of our approach to linear equalizers, Wiener filtering, and decoding for intersymbol interference channels is also discussed in this section. In Section IX, the performance of the system is evaluated. A glossary of terms appears at the end of the article.

II. System Description

The received signal plus noise is denoted by $r(t) = s(t - \tau_0) + n(t)$, where $s(t)$ is the signal, $n(t)$ is AWGN and τ_0 is the delay from the transmitter to the receiver. The transmitted signal $s(t)$ is

$$s(t) = \sum_k a_k \, p(t - kT) \tag{1}$$

a sequence of pulses with a pulse-shaping waveform $p(t)$. The input alphabet U is a finite alphabet with $a_i \epsilon U = \{\pm 1, \pm 2, \ldots, \}$.

The analog IDF is shown in Fig. 1(a). The analog IDF is an ideal matched filter when $p(t)$ is a rectangular pulse from $t = 0$ to $t = T$. It detects the kth symbol by integrating over time $kT + \tau_0$ to $(k + 1)T + \tau_0$.

The digital IDF is depicted in Fig. 1(b). In the digital implementation a low-pass anti-aliasing filter is used for filtering the input signal. In this article the one-sided bandwidth of this filter is denoted by W(Hz). The filter output is sampled, with the ith sample occurring at time $iT_s + \tau_1$. The digital IDF detects the kth symbol by summing all the samples from $t = kT + \tau_0$ to $t = (k + 1)T + \tau_0$.

We assume that there is perfect symbol synchronization at the receiver, in the sense that the beginning and end times of each symbol are known. For the kth symbol the "Sampling Offset" is defined by the length of time after the start of the symbol to when the first sample in the symbol occurs. This time is $(iT_s + \tau_1) - (kT + \tau_0)$ for the smallest i such that the expression is nonnegative. The first sample of each symbol may occur anywhere between 0 and T_s seconds after the beginning of the symbol. A typical symbol waveform and the sampling points are shown in Fig. 2.

To illustrate the effect of offset in sampling, Fig. 3 depicts one pulse of the sampled waveform for an alternating rectangular data pattern of length 21, when the anti-aliasing filter is an ideal low-pass filter. The sampled waveform for the 11th symbol, a -1 pulse, is shown in Fig. 3 for $WT = 2$, and for $T = 4T_s$. The filtered waveform is not rectangular due to the finite bandwidth of the anti-aliasing filter. In Fig. 3, for every sampling offset value, with increments $T_s * 0.05$, a unique English letter (a through t) is used to indicate the point at which the sample occurs. Every letter occurs four times, corresponding to the four samples per symbol.

In an earlier article [1], we considered the effect of offset in the digital IDF. It was shown that the loss due to offset in sampling is significant when the number of samples per symbol is low ($T/T_s < 8$). The loss depends on the bandwidth W, the sampling rate, and the relative phase of the samples and symbols. If the signal is sampled at the optimum sampling time the loss is relatively small. This loss is due to bandwidth limiting of the input signal. For example, when four samples per symbol are used in the digital IDF, i.e., $T/T_s = 4$ and $WT = 2$, the worst case loss is approximately 1.2 dB averaged over random data patterns. This occurs when the offset is zero, indicated by the letter a in Fig. 3. The minimum loss is 0.35 dB when the offset is $T_s/2$, indicated by the letter j in Fig. 3. Thus a variation of 0.8 dB in the loss occurs due to the phase of the offset in sampling. To decrease this variation and, as a result, to reduce system sensitivity due to the offset, we are led to consider the Weighted Integrate-and-Dump Filter (WIDF), which is the main subject of this article.

III. Derivation of the WIDF Using MSE

In Fig. 1 the IDF is shown for both the analog and digital implementations. The sampled output of the analog IDF is $A(kT)$, denoted as simply A_k. In this section, the minimum mean square error criterion is used to estimate the sequence A_k from the digital samples.

We formulate the problem in the context of Fig. 4(a). In this figure the digital IDF filter is denoted by $f(\cdot)$. The operator $f(\cdot)$ maps the observation vector $\mathbf{y} = (y_1, y_2, \cdots, y_N)$ in the kth symbol onto $\hat{A}(kT)$, an estimate of $A(kT)$.

We seek to find $f(\cdot)$ such that the minimum mean square error criterion is minimized, i.e., we minimize

$$E[(A_k - \hat{A}_k)^2 \mid \mathbf{y}] \tag{2}$$

where $\hat{A}_k = f(\mathbf{y})$, and $E[\cdot]$ denotes the expectation operator. Note here that the estimate of A_k is based only on the observation vector during a single symbol time. In Section III, we

briefly discuss the case when this restriction is relaxed, when the relationship of WIDF to linear equalizers is pointed out.

It can be shown [2]-[4] that the optimal $f(\mathbf{y})$ is the conditional expectation of A_k conditioned on the observed vector $\mathbf{y}$:

$$f(\mathbf{y}) = E[A_k \mid \mathbf{y}] \tag{3}$$

Since $r(t) = s(t) + n(t)$, and $n(t)$ is AWGN, the conditional probability density function of $r(t)$ conditioned on the input data sequence $\mathbf{a}$ is Gaussian. However, the conditional probability density function of A_k conditioned on $\mathbf{y}$ is not Gaussian due to Inter Symbol Interference (ISI), and it is almost impossible to explicitly evaluate this probability density function. We assume this density function is Gaussian, and hence the conditional expectation is a linear function of the observed vector $\mathbf{y}$. Thus, under this assumption,

$$E[A_k \mid \mathbf{y}] = \sum_{i=1}^{N} w_i y_i \tag{4}$$

We shall not state the complete derivation of the Linear Minimum Mean Square Error (LMMSE) criterion. Interested readers could refer to [2], [3] to obtain the complete derivation of the following result.

The optimum weight sequence $\mathbf{w} = (w_1, w_2, \ldots, w_N)$ may be expressed in terms of the second order statistics of the observed vector $\mathbf{y}$ as

$$\mathbf{w} = \mathbf{R}_{\mathbf{yy}}^{-1} \mathbf{R}_{\mathbf{yA}} \tag{5}$$

where the matrix $\mathbf{R}_{\mathbf{yy}}$ is the autocorrelation matrix (assuming $\mathbf{R}_{\mathbf{yy}}$ is nonsingular) with elements $E[y_i y_i]$, and $\mathbf{R}_{\mathbf{yA}}$ is the cross-correlation vector between y_i and A_k, with elements $E[y_i A_k]$, where $E[\cdot]$ denotes the expectation operator.

$$\mathbf{R}_{\mathbf{yy}} = \begin{bmatrix} E[y_1 y_1] & E[y_1 y_2] & \cdots & E[y_1 y_N] \\ E[y_2 y_1] & E[y_2 y_2] & & \\ \vdots & & \ddots & \\ E[y_N y_1] & & & E[y_N y_N] \end{bmatrix} \tag{6}$$

and

$$\mathbf{R}_{\mathbf{yA}} = \begin{bmatrix} E[y_1 A_k] \\ E[y_2 A_k] \\ \vdots \\ E[y_N A_k] \end{bmatrix} \tag{7}$$

In order to evaluate the matrix $\mathbf{R}_{\mathbf{yy}}$ and the vector $\mathbf{R}_{\mathbf{yA}}$, in the following section a linear system is specified which generates the autocorrelation functions $R_{yy}(\tau)$ and $R_{yA}(\tau)$. The matrix $\mathbf{R}_{\mathbf{yy}}$ and the vector $\mathbf{R}_{\mathbf{yA}}$ are obtained by sampling the autocorrelation functions at time $t = iT_s + \delta$, for $i \in [1, N]$, where δ is the offset.

IV. Evaluation of Matrix $\mathbf{R}_{\mathbf{yy}}$ and Vector $\mathbf{R}_{\mathbf{yA}}$

The following results are a direct consequence of the application of second order statistics of a stationary stochastic process to the input-output relations of a linear system [3], [5]. Throughout this article on-line "$*$" denotes convolution, and superscript "$*$" denotes the complex conjugate.

To compute the autocorrelation function $R_{yy}(\tau)$, note that

$$y(t) = s(t) * h(t) + n(t) * h(t) \tag{8}$$

let $x(t) = s(t) * h(t)$ represent the filtered signal component, and $z(t) = n(t) * h(t)$ represent the filtered noise component of $y(t)$. The autocorrelation function $R_{yy}(\tau)$ may be expressed in terms of the cross-correlation of $R_{ys}(\tau)$ and $R_{yn}(\tau)$ as

$$R_{yy}(\tau) = R_{ys}(\tau) * h(\tau) + R_{yn}(\tau) * h(\tau) \tag{9}$$

The two cross-correlation functions $R_{ys}(\tau)$ and $R_{yn}(\tau)$ are

$$\begin{aligned} R_{ys}(\tau) &= R_{ss}(\tau) * h^*(-\tau) \\ R_{yn}(\tau) &= R_{nn}(\tau) * h^*(-\tau) \end{aligned} \tag{10}$$

To compute $R_{yA}(\tau)$, note that

$$A_0 = \int_0^T s(\xi) d\xi \tag{11}$$

The cross-correlation function $R_{yA}(\tau)$ may be expressed as

$$R_{yA}(\tau) = E[y(\tau)A(T)]$$

$$= -\int_{\tau}^{\tau-T} \left(R_{ys}(\xi) + R_{ny}(\xi)\right) d\xi \qquad (12)$$

for a fixed T, where the pair $R_{ys}(\tau)$ and $R_{yn}(\tau)$ are given by (10).

Figure 5 depicts a linear system which can be used to evaluate the two autocorrelation functions R_{yy} and R_{yA}. In Fig. 5 the input to the system is the autocorrelation function of the received signal, and the outputs are the desired autocorrelation functions $R_{yy}(\tau)$ and $R_{yA}(\tau)$. The sampled sequence of $R_{yy}(\tau)$ and $R_{yA}(\tau)$ generates the corresponding matrices $\mathbf{R_{yy}}$ and $\mathbf{R_{yA}}$. This illustrates a method to obtain the correlation necessary to calculate the optimal weight sequence w, from (5).

V. Average Signal Response

We now seek to determine the average signal response for the output of the WIDF.

The response of the low pass filter to the observed signal $r(t)$ is

$$y(t) = \int_{-\infty}^{\infty} h(t-\xi)\, s(\xi - \tau_0)\, d\xi$$

$$+ \int_{-\infty}^{\infty} h(t-\xi)\, n(\xi)\, d\xi \qquad (13)$$

Using (1) for $s(t)$ we have

$$y(t) = \sum_{k'=-\infty}^{\infty} \int_{-\infty}^{\infty} a_{k'}\, h(t-\xi)\, p(\xi - k'T - \tau_0)\, d\xi$$

$$+ \int_{-\infty}^{\infty} h(t-\xi)\, n(\xi)\, d\xi \qquad (14)$$

The signal $y(t)$ is sampled each T_s sec, at time $iT_s + \tau_1$. We denote $y(iT_s + \tau_1)$ as y_i. Taking the expectation of (14) conditioned on a given data sequence **a** and noting that the noise $n(t)$ is assumed to have zero mean, the conditional expectation of y_i is

$$E[y_i \mid \mathbf{a}] = \sum_{k'} a_{k'} \int_{-\infty}^{\infty} h(iT_s + \tau_1 - \xi)\, p(\xi - kT - \tau_0) d\xi \qquad (15)$$

With a change of variable (15) can be written as

$$E[y_i \mid \mathbf{a}] = \sum_{k'} a_{k'} \int_{-\infty}^{\infty} h(iT_s - kT + \delta - x)\, p(x) dx \qquad (16)$$

where $\delta = \tau_1 - \tau_0$. Let

$$q_i(k, \delta) = \int_{-\infty}^{\infty} h(iT_s - kT + \delta - x)\, p(x) dx \qquad (17)$$

represent the signal response of the filter at time $iT_s + \tau_1$ due to a single pulse at time $kT + \tau_0$. For simplicity we denote $q_i(k, \delta)$ as simply $q_i(k)$. The total average signal response from (16), for a given fixed δ, may be expressed as

$$E[y_i \mid \mathbf{a}, \delta] = \sum_{k'} a_{k'}\, q_i(k') \qquad (18)$$

Let I^k be the set of all i such that the ith sample falls in the kth symbol time, i.e.,

$$I^k = \{i:\ kT \leqslant iT_s + \delta < (k+1)T\} \qquad (19)$$

The WIDF output for the kth symbol, denoted by A_k, is

$$\hat{A}_k = \sum_{i \in I^k} y_i\, w_i \qquad (20)$$

The expectation of $\hat{A}_k$ over the noise, conditioned on **a** and δ, is

$$E[\hat{A}_k \mid \mathbf{a}, \delta] = \sum_{i \in I^k} \sum_{k'} a_{k'}\, q_i(k')\, w_i \qquad (21)$$

To further simplify this expression, define the event indicator function which is 1 if and only if $i \epsilon I^k$, i.e.,

$$\ell_i(\delta, k) = \begin{cases} 1 & \text{when } kT \leq iT_s + \delta < (k+1)T \\ 0 & \text{otherwise} \end{cases} \quad (22)$$

Thus from (21) we have

$$E[\hat{A}_k \mid \mathbf{a}; \delta] = \sum_i \sum_{k'} a_{k'} \ell_i(\delta, k) w_i q_i(k') \quad (23)$$

VI. Noise Response

Now we consider the noise response of WIDF in order to compute the total SNR at the output of the WIDF. Let z_i denote the sampled noise response of the filter at time $iT_s + \tau_1$.

$$z_i = \int_{-\infty}^{\infty} n(\xi) h(iT_s + \tau_1 - \xi)\, d\xi \quad (24)$$

Since the WIDF is a linear system, the variance of $\hat{A}_k$ conditioned on $\mathbf{a}$ is equal to the variance of the response of the kth symbol due to noise alone, i.e., it is independent of $s(t)$. The variance of $\hat{A}_k$ is

$$\text{var}[\hat{A}_k \mid \delta, \mathbf{a}] = \sum_{i \epsilon I^k} \sum_{j \epsilon I^k} E[z_i z_j]\, w_i w_j \quad (25)$$

Note that this variance does depend on δ and k, since the number of samples occurring in the kth symbol depends on δ. Using (22) and noting that $E[n(t)n(\tau)] = N_0/2\, \delta_0(t - \tau)$ (δ_0 here is the Dirac delta function), we have

$$\text{var}[\hat{A}_k \mid \delta, \mathbf{a}] = \sum_i \sum_j \ell_i(k, \delta) \ell_j(k, \delta) w_i w_j R_z\left((i-j)T_s\right)$$

$$R_z[(i-j)T_s] = \frac{N_0}{2} \int_{-\infty}^{\infty} h\left((i-j)T_s - \xi\right) h(\xi) d\xi \quad (26)$$

where $R_z(\cdot)$ is the autocorrelation of z_i.

VII. Definition of SNR Loss

In this section, we define a measure to evaluate the degradation which results in using the WIDF as opposed to analog IDF. The analog IDF of Fig. 1(a) is the optimum matched filter when $p(t) = 1$ for $0 < t < T$ and zero otherwise. We define SNR at the IDF output as the ratio of the square of the mean to the variance. Denoting SNR_A for the analog IDF, it is well known [6] that

$$\text{SNR}_A = \frac{2A^2 T}{N_0} \quad (27)$$

We assume with no loss of generality that the signal amplitude $A = 1$. Denoting SNR_D as the SNR at the output of the WIDF, we compare the SNR_D with the analog IDF by considering the ratio

$$\gamma \triangleq \frac{\text{SNR}_D}{\text{SNR}_A} \quad (28)$$

Define

$$\text{SNR}_D \triangleq \frac{(E[\hat{A}_k \mid \mathbf{a}, \delta])^2}{\text{var}[\hat{A}_k \mid \mathbf{a}, \delta]}$$

and then we have

$$\gamma = \frac{N_0}{2T} \text{SNR}_D \quad (29)$$

In the remaining sections $\gamma_{\text{dB}} = 10^*\log_{10}(\gamma)$ (dB) is computed for various filters and data patterns. Normally $\gamma < 1$, because the digital IDF has a loss with respect to the analog IDF. The loss in dB is $-\gamma_{\text{dB}}$. The minimum loss corresponds to the maximum γ which typically approaches one ($\gamma_{\text{dB}} = 0$ dB). Maximum loss is unbounded and corresponds to infinity (in dB).

VIII. WIDF for Rectangular Pulse and Ideal Filter

In general, the pulse shape $p(t)$ may be chosen to take numerous shapes (e.g., raised root-cosine). In some cases, it is chosen to extend over more than one symbol duration, such as for partial response signaling (sometimes referred to as correlated coding or controlled intersymbol interference). For bandwidth-limited channels, the pulse shape and duration are selected to increase the bandwidth efficiency of the communication system.

The motivation to consider the ideal low pass filter is to eliminate aliasing in an ideal manner. The use of a realizable filter such as Butherworth or Chebyshev [7] does not greatly influence the results, since the realizable filter can be consid-

ered as an approximation to the unrealizable filter with finite group delay [7].

We consider only non-overlapping rectangular pulses throughout the rest of this article, since this pulse shape has traditionally been used for NASA's deep space missions.

In the case of the rectangular pulse we simply have

$$p(t) = \begin{cases} 1 & t \in [0,T] \\ 0 & \text{otherwise} \end{cases}$$

Then from (17), $q_i(k)$ is

$$q_i(k) = \int_0^T h(iT_s - kT + \delta - x)\,dx \tag{30}$$

and from (21), the average signal response is

$$E[\hat{A}_k \mid \mathbf{a}, \delta] = \sum_{k'} \sum_{i \in I^k} a_{k'} q_i(k') w_i \tag{31}$$

The ideal low pass filter with unit gain and low pass bandwidth W – Hz is noncausal with impulse response

$$h(t) = 2W \frac{\sin 2\pi W t}{2\pi W t} = 2W \sin c(2\pi W t) \tag{32}$$

The expression for the signal response in (30) does not evaluate to a closed form in this case, but is

$$q_i(k) = \frac{1}{\pi} \int_0^T \frac{\sin 2\pi W(iT_s - kt + \delta - x)}{(iT_s - kT + \delta - x)}\,dx \tag{33}$$

It is possible to express (33) in terms of

$$Si(x) = \int_0^x \frac{\sin u}{u}\,du$$

as

$$q_i(k) = \frac{1}{\pi}\,[Si(2\pi W(iT_s - (k+1)T + \delta)) - Si(2\pi W(iT_s - kT + \delta))] \tag{34}$$

Inserting (34) into (31) yields the average signal response. To find the noise variance, it suffices to note that the noise spectral density at the output of the filter is

$$S_n(f) = \begin{cases} \frac{N_0}{2} & |f| < W \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

and thus the autocorrelation function is

$$R_n(\tau) = \frac{N_0 W}{2} \frac{\sin 2\pi W \tau}{2\pi W \tau} \tag{36}$$

Thus, the noise variance at the output of WIDF can be expressed from (26) and (36) as

$$\text{var}[\hat{A}_k \mid \delta, \mathbf{a}] = \frac{N_0 W}{2} \sum_i \sum_j w_i \ell_i(k, \delta)\, w_j \ell_j(k, \delta) \sin c(2\pi(i-j)2WT_s) \tag{37}$$

Thus from (29) and (37), γ can be evaluated for arbitrary $\mathbf{w}$ and rectangular pulse shapes and ideal filters as

$$\gamma = \frac{\left[\frac{1}{\pi} \sum_{i \in I^k} \sum_{k'} a_{k'} w_i \left(Si(2\pi W(iT_s - (k+1)T + \delta)) - Si(2\pi W(iT_s - kT + \delta))\right)\right]^2}{WT \sum_{i \in I^k} \sum_{j \in I^k} w_i w_j \sin|c(2\pi W(i-j)T_s)} \tag{38}$$

The next step is to compute the optimum weights according to (5). This requires the evaluation of R_{yy} and R_{yA} given by (9) and (12). We consider the special case in this section where the filter $h(t)$ is an ideal rectangular filter and $p(t)$ is a rectangular pulse. Referring to Fig. 5, and using the $h^*(-\tau) = h(\tau)$, one can verify that for an ideal filter

$$R_{yy}(\tau) = R_{ss}(\tau) * h(\tau) + R_{nn}(\tau) * h(\tau) \tag{39}$$

The signal autocorrelation function $R_{ss}(\tau)$ for a random binary waveform [7] is

$$R_{ss}(\tau) = \begin{cases} 1 - \frac{|\tau|}{T} & |\tau| < T \\ 0 & \text{otherwise} \end{cases} \tag{40}$$

The ideal filter impulse response as in (32) is

$$h(t) = 2W \sin c(2\pi W t) \tag{41}$$

For simplicity, let $R_x(\tau) = R_{ss}(\tau) * h(\tau)$ and $R_z(\tau) = R_{nn}(\tau) * h(\tau)$. Thus $R_x(\tau)$ is

$$R_x(\tau) = \int_{-\infty}^{\infty} R_s(\beta)\, h(\tau - \beta) d\beta \tag{42}$$

Since the input signal autocorrelation function (40) is nonzero only in the interval $[-T, T]$, (42) is

$$R_x(\tau) = \int_{-T}^{T} R_{ss}(\beta)\, h(\tau - \beta) d\beta \tag{43}$$

This integral can be explicitly evaluated for $h(t)$ in (41) by decomposing it into two successive integrals

$$R_x(\tau) = \int_{-T}^{0} R_{ss}(\beta)\, h(\tau - \beta) d\beta + \int_{0}^{T} R_{ss}(\beta)\, h(\tau - \beta) d\beta \tag{44}$$

After some manipulation (44) can be explicitly evaluated in terms of the $Si(\cdot)$ function, and it yields

$$\begin{aligned} R_x(\tau) = \frac{1}{\pi} \Bigg\{ & \frac{\tau}{T} \left[Si(\mathrm{B}(\tau - T)) + Si(\mathrm{B}(\tau + T)) - 2\, Si(\mathrm{B}\tau) \right] \\ & + Si(\mathrm{B}(\tau - T)) - Si(\mathrm{B}(\tau - T)) \\ & + \frac{1}{\mathrm{B}T} \left[\cos(\mathrm{B}(\tau + T)) + \cos(\mathrm{B}(\tau - T)) - 2 \cos(\mathrm{B}\tau) \right] \Bigg\} \end{aligned} \tag{45}$$

where $\mathrm{B} = 2\pi W$.

To evaluate $R_z(\tau)$, consider

$$R_z(\tau) = \int_{-\infty}^{\infty} R_n(\xi)\, h(\tau - \xi)\, d\xi \tag{46}$$

which is simply

$$R_z(\tau) = \frac{W N_0}{2} \sin c(2\pi W \tau) \tag{47}$$

To evaluate $R_{yA}(\tau)$ we need to integrate (45) and (47) over $[t - T, t]$. The expression in (45) does not evaluate to a closed form expression, but integrating (47) over this interval yields

$$\int_{t-T}^{t} R_z(\tau) = \frac{N_0 W}{2} \left(-Si(\mathrm{B}(t - T)) + Si(\mathrm{B}\tau) \right) \tag{48}$$

Thus, we have

$$R_{yA} = \int_{\tau-T}^{\tau} R_x(t) dt + \frac{N_0 W}{2} \left(Si(\mathrm{B}\tau) \right) - Si(\mathrm{B}(\tau - T)) \tag{49}$$

and $R_{yy}(\tau)$ is

$$R_{yy}(\tau) = R_x(\tau) + R_z(\tau) \tag{50}$$

where $R_x(\tau)$ is defined in (45) and $R_z(\tau)$ is defined in (47). The optimum weights are calculated using (49) and (50) in (5).

A. Relationship to Linear Equalizer

In general, the signal processing algorithm that is designed to compensate for the ISI of the communication channel is referred to as an "equalizer." The most common method for equalization is a transversal filter [8], which is designed such that its coefficients optimize the performance of a system according to criteria selected by the designer.

When the MSE criterion is used to obtain the tap weight coefficients of the equalizer, the equalizer is equivalent to the WIDF when N, the length of the observation vector y in (2), exceeds the number samples in a single symbol time, i.e., $N > T/T_s$. It is pointed out that all our results will hold in this case, and our analysis for derivation of the WIDF can be effectively used for designing LMSE equalizers.

The optimal decoding algorithm for channels with ISI uses the maximum likelihood sequence estimation. Viterbi and Omura [9] discuss optimal decoding for ISI channels using the maximum likelihood sequence estimation, and they formulate the application of the Viterbi algorithm for estimating the data sequence, which results in a nonlinear estimator.

B. Relationship to Wiener and Kalman Filtering

If the length of the observation vector y in (2) is infinite ($N = \infty$), it is well known that the optimal matched filter is the discrete time Wiener filter [2]. The Wiener–Hopf method requires the factorization of a spectral density matrix. Analytical solutions for this method are very difficult to derive, and

even when they do exist, it is an arduous task to physically realize such filters.

For lumped processes [2] which result by passing the received signal through a realizable filter, it is possible to model the observation process using a state space model. In this case Kalman filtering [2]-[4] can be applied to both vector observation (finite N) and time varying state space models. That subject is beyond the scope of this article.

IX. Performance Analysis

In this section we compute the set of weight coefficients for the case when $N = 4$, evaluate the degradation of the WIDF, and compare its loss with the digital IDF.

The software simulation programs explicitly compute (38) for arbitrary input signal sequences **a**, when an ideal filter is used and the input pulse shape $p(t)$ is a rectangular pulse.

In Table 1, the optimum set of weight coefficients for the case when $T/T_s = 4$ and $WT = 2$ is shown. These weights were computed using (5) and computing R_{yy} and R_{yA} using (49) and (50).

The output of the ideal low pass filter depends on both the past and future inputs. To approximate this, we consider a 21-symbol block, and the 11th symbol is analyzed for each data pattern. A block of 21 symbols was found to be sufficiently long to analyze the IDF [1].

Figure 6 shows the performance when the input data pattern **a** is a sequence of alternating +1, −1 sequences, i.e., **a** = (+1, −1, +1, −1, . . .). The degradation for the 11th symbol is shown for both the IDF and WIDF. The WIDF is less lossy than the IDF for all values of offset. For the best offset, 0.5, the loss is 0.4 dB for the WIDF and 0.46 dB for the IDF, a minimal difference. This loss is mainly due to the bandwidth limiting. For the worst case offset, 0.0, the WIDF is more than 1.0 dB better than the IDF, with a loss of 0.77 dB for the WIDF and 1.81 dB for the IDF. The variation in the degradation due to offset decreases from 1.35 dB to only 0.33 dB for the WIDF.

In Fig. 7, the average loss is shown for a random binary vector of 4640 symbols, using the same set of weighing coefficients. For each offset, the average is computed by breaking the 4640 symbol vector into 220 blocks 21 symbols long and computing the loss of the 11th symbol for each block, and finally computing the average loss. For the best offset, the losses for the WIDF and IDF are again similar, and they are slightly less than for the alternating data pattern. The worst case losses, averaged over the data patterns, are 1.26 dB for the IDF but only 0.68 dB for the WIDF, an improvement of 0.58 dB. For the WIDF, the variation in average performance over offset is less than 0.3 dB.

X. Conclusion

In this article, based on the MSE criterion, a new class of digital matched filters was derived, which approximates the analog IDF. A linear system was outlined that generates the autocorrelation functions necessary to evaluate the WIDF. The SNR loss due to using WIDF was formulated.

The WIDF weighting coefficients which are optimum in the mean square sense were computed for the case when an ideal filter and rectangular pulse shape are used. The performance for this case was evaluated for the case of four samples per symbol. It was shown that the variation due to offset was reduced to under 0.3 dB for the WIDF, from almost 1 dB for the IDF, averaged over a random pattern. Compared to a system with the samples phase locked to the symbol clock, a WIDF with offset sampling suffers a worst case offset loss of less than 0.3 dB, and an average loss of less than 0.15 dB.

This improved performance means that lower sampling and processing rates can be used for a given symbol rate, resulting in reduced system cost. Alternately, for a fixed bandwidth and sampling rate, higher telemetry rates are enabled. Telemetry symbol rates of one-half the bandwidth and one-fourth of the sampling rate can be realized with a loss due to bandwidth limiting, sampling and filtering of less than 0.6 dB. An unweighted IDF [1] would require approximately twice the bandwidth and twice the sampling rate for the same performance.

Glossary of Terms

T_s	Sampling time in sec
T	Symbol time in sec
$2W$	Filter bandwidth in hertz
a_i	Transmitted symbol
$p(t)$	Pulse shaping waveform
A_k	The sampled output of Integrate-and-Dump Filter at time k
y_i	The sampled output of the prefilter
$s(t)$	The transmitted signal
$n(t)$	Additive White Gaussian Noise with flat spectral density $N_0/2$
$r(t)$	The received signal
$y(t)$	The output of the low-pass prefilter
$h(t)$	The transfer function of the low-pass prefilter
τ_0	Transport lag from transmitter to receiver in sec
τ_1	Sampling offset in sec
$R_x(\tau)$	Autocorrelation function of signal $x(t)$

References

[1] R. Sadr and W. J. Hurd, "Detection of Signals by the Digital Integrate-and-Dump Filter with Offset Sampling," *TDA Progress Report 42-91*, vol. July–September 1987, Jet Propulsion Laboratory, Pasadena, California, November 15, 1987.

[2] T. Kailath, *Lectures on Linear Least-Squares Estimation*, Amsterdam: Springer Verlag, 1976.

[3] A. V. Balakrishnan, *Stochastic Filtering and Control*, Los Angeles: Optimization Software, 1981.

[4] A. H. Jaswinsky, *Stochastic Processes and Filtering Theory*, New York: Academic Press, 1970.

[5] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, New York: McGraw-Hill, 1965.

[6] D. A. Whalen, *Detection of Signals in Noise*, New York: Academic Press, 1971.

[7] A. Papoulis, *Signal Analysis*, New York: McGraw-Hill, 1977.

[8] L. M. Honig and D. G. Messerschmidt, *Adaptive Filter*, Norwell, Massachusetts: Kluwer Academic Publishers, 1984.

[9] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, New York: McGraw-Hill, 1979.

Table 1. Weight coefficients for WIDF $N = 4$, $WT = 2$

Offset	w_1	w_2	w_3	w_4
0	0.5	1.24	0.9	1.24
0.05 T_s	0.5	1.22	0.9	1.22
0.1 T_s	0.6	1.20	0.9	1.20
0.15 T_s	0.6	1.18	0.9	1.18
0.2 T_s	0.7	1.16	0.9	1.16
0.25 T_s	0.7	1.13	0.9	1.13
0.3 T_s	0.8	1.12	0.9	1.10
0.35 T_s	0.8	1.09	0.9	1.07
0.4 T_s	0.8	1.07	1.01	1.03
0.45 T_s	0.9	1.05	1.03	1.01
0.5 T_s	0.9	1.03	1.03	0.9
0.55 T_s	1.01	1.03	1.05	0.9
0.6 T_s	1.03	1.01	1.07	0.8
0.65 T_s	1.07	0.9	1.09	0.8
0.7 T_s	1.10	0.9	1.12	0.8
0.75 T_s	1.13	0.9	1.13	0.7
0.8 T_s	1.16	0.9	1.16	0.7
0.85 T_s	1.18	0.9	1.18	0.6
0.9 T_s	1.20	0.9	1.20	0.6
0.95 T_s	1.22	0.9	1.22	0.5
T_s	1.24	0.9	1.24	0.5

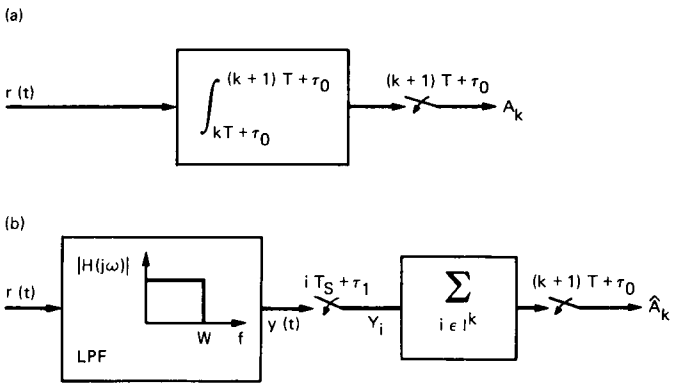


Fig. 1. Integrate-and-dump filters: (a) analog; (b) digital

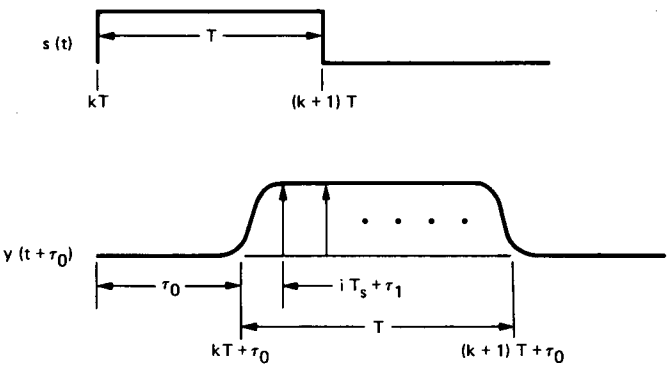


Fig. 2. Offset in sampling

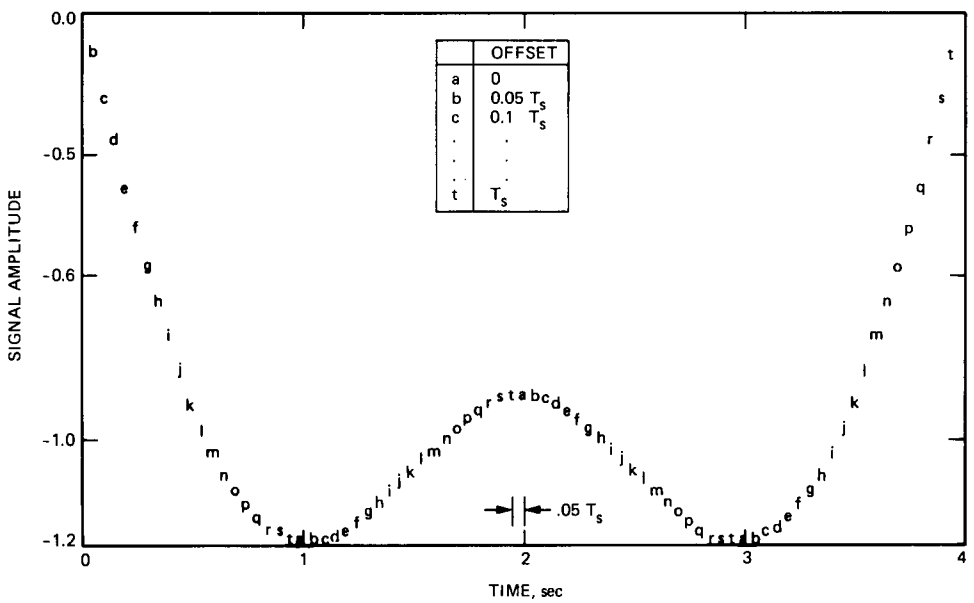


Fig. 3. Sampled waveform ($WT = 2$), alternating data pattern with $T/T_s = 4$

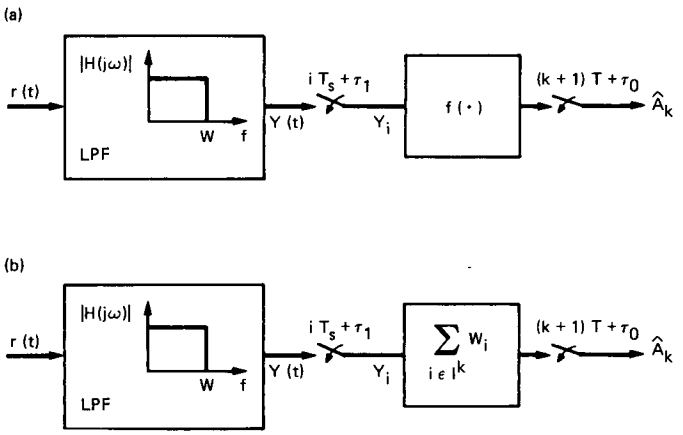


Fig. 4. Digital matched filtering: (a) optimum digital matched filter; (b) weighted integrate-and-dump filter

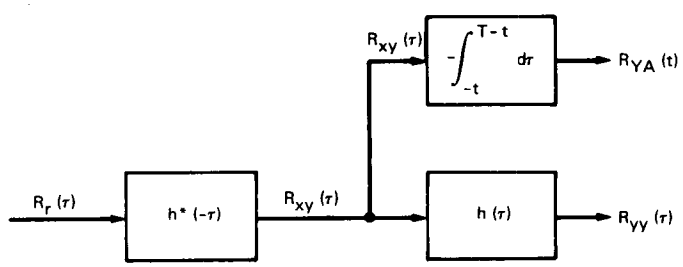


Fig. 5. Linear system to generate $R_{yy}(\tau)$, $R_{yA}(\tau)$

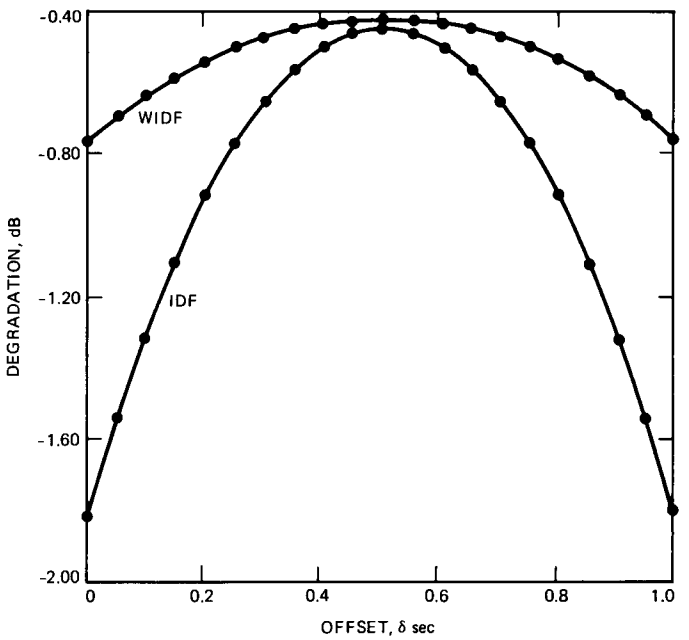


Fig. 6. Comparison of λ_{dB} for IDF and WIDF for alternating data pattern

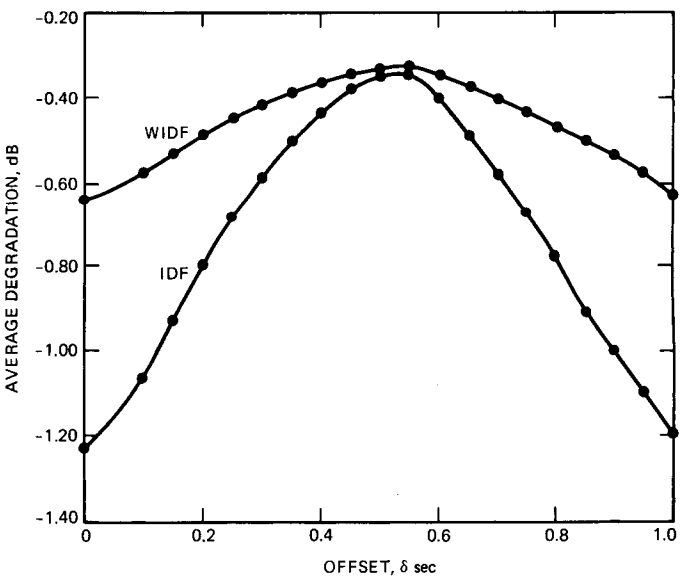


Fig. 7. Comparison of λ_{dB} for IDF and WIDF for random data pattern

The Design Plan of a VLSI Single Chip (255, 223) Reed-Solomon Decoder

I. S. Hsu, H. M. Shao, and L. J. Deutsch
Communications Systems Research Section

The VLSI architecture of a single chip (255, 223) Reed-Solomon decoder for decoding both errors and erasures is described in this article. A decoding failure detection capability is also included in this system so that the decoder will recognize a failure to decode instead of introducing additional errors. This could happen whenever the received word contains too many errors and erasures for the code to correct. The number of transistors needed to implement this decoder is estimated at about 75,000 if the delay for received message is not included. This is in contrast to the older transform decoding algorithm which needs about 100,000 transistors. However, the transform decoder is simpler in architecture than the time decoder. It is therefore possible to implement a single chip (255, 223) Reed-Solomon decoder with today's VLSI technology. An implementation strategy for the decoder system is presented. This represents the first step in a plan to take advantage of advanced coding techniques to realize a 2.0-dB coding gain for future space missions.

I. Introduction

A concatenated coding system consisting of a convolutional inner code and a Reed-Solomon outer code has been adopted as a guideline for downlink telemetry for future space missions by CCSDS (Consultative Committee for Space Data Systems) (Ref. 1). The convolutional inner code is the same (7, 1/2) code used by NASA's Voyager project. The outer Reed-Solomon code is a (255, 223) block code of 8-bit symbols and it is capable of correcting t errors and s erasures if $2t + s \leqslant 32$. The performance of such schemes is investigated in Ref. 2 where it is shown that this concatenated channel provides a coding gain of almost 2 dB over the convolutional-only channel at a decoded bit error rate of 10^{-5} using only the error correcting capability of the codes. In Ref. 3, a (255, 223) Reed-Solomon encoder using Berlekamp's bit-serial multiplication algorithm is developed and proved to perform well. However, due to the sophisticated procedures involved in the Reed-Solomon decoding algorithm, especially the portion to perform the Euclid's algorithm, it is much more difficult to design a Reed-Solomon decoder in VLSI.

Recently, Brent and Kung (Ref. 4) suggested a systolic array architecture to compute the greatest common divisor (GCD) of two polynomials. By the use of this idea, a VLSI design of a pipeline Reed-Solomon decoder was developed (Ref. 6). Hsu et al. used this idea to implement a VLSI chip for calculating the GCD of two polynomials with 8-bit coefficients (Ref. 5). More recently, the pipeline design of the Reed-Solomon decoder was revised in Ref. 5. In this revision

the transform decoding algorithm is replaced by a time domain algorithm (see Appendix B). Erasure correction capability is included in the revision, and a multiplexed design for the Euclid's algorithm is used instead of the systolic array design. These improvements reduce the circuitry needed for VLSI implementation of Reed-Solomon decoder.

In this article, the VLSI architecture of the (255, 223) Reed-Solomon decoder for decoding both errors and erasures is described. The functional behavior of each block will be explained and the number of transistors needed for VLSI implementation is estimated. Finally, comparisons with the previous design are included as well.

This article represents a first cut at a final implementation plan for a Reed-Solomon decoding subsystem that may be used to support future deep space missions within the Deep Space Network (DSN). It also constitutes the first step toward the realization of a 2.0-dB coding gain through the use of advanced error correcting coding strategies.

II. The Decoding Procedure

Let $N = 2^m - 1$ be the length of a (N, I) Reed-Solomon code over $GF(2^m)$ with design distance d. Suppose that t errors and s erasures occur, and $s + 2t \leqslant d$. Then Reed-Solomon coding theory implies that the original codeword may be recovered from the received data.

First some definitions are needed. Let each X_i be an error location or an erasure location, and define the two sets $\Lambda = \{X_i | X_i$ is an erasure location$\}$, and $\lambda = \{X_i | X_i$ is an error location$\}$. Let Y_i be the corresponding errata magnitude and let $r = (r_0, r_1, \ldots, r_{N-1})$ be the received vector. Now the decoding process may be described in terms of the following basic steps.

Step 1: Compute the syndrome polynomial

$$S(Z) = \sum_{k=1}^{\infty} S_k Z^{-k}$$

where

$$S_k = \sum_{n=0}^{N-1} r_n \alpha^{nk}, \qquad 1 \leqslant k \leqslant d-1 \tag{1}$$

The numbers S_k are called the "syndromes" of the received word.

Step 2: Compute the erasure locator polynomial

$$\Lambda(Z) = \prod_{X_i \in \Lambda} (Z - X_i) \tag{2}$$

Step 3: Multiply $S(Z)$ and $\Lambda(Z)$ to obtain the Forney syndrome polynomial

$$T(Z) = S(Z)\,\Lambda(Z) \tag{3}$$

Step 4: Compute the errata evaluator polynomial $A(Z)$ and the error locator polynomial $\Lambda(Z)$ from

$$T(Z) = \frac{A(Z)}{\Lambda(Z)}$$

by a modified Euclid's algorithm.

Step 5: Multiply $\lambda(Z)$ and $\Lambda(Z)$ to get the errata locator polynomial

$$P(Z) = \lambda(Z)\,\Lambda(Z) \tag{4}$$

Step 6: Perform a Chien search on $P(Z)$ to find the error location set λ and the erasure location set Λ.

Step 7: Compute the errata magnitudes

$$Y_k = \frac{A(X_k)}{X_k P'(X_k)}, \qquad 1 \leqslant k \leqslant s + t$$

by evaluating $A(Z)$ and $P'(Z)$, the derivative of $P(Z)$. Use the sets Λ and λ to direct the addition of Y_k to the received vector r to produce the decoded result.

The extra calculation required to recognize a failure to decode was left out of the above discussion for clarity. It is explained in Ref. 7.

III. Functional Description and Transistor Estimations

In this section, the decoder is broken down into several basic blocks. The function of each block is described, and the number of transistors needed for VLSI implementation is estimated. The discussion here will be more detailed than in Ref. 7.

The estimates of the number of transistors required for the various blocks are based on the considerable work that has already been done on this project and the combined expertise

of a team of logic and VLSI designers within the Digital Signal Processing Research Group.

Figure 1 shows the overall architecture of the VLSI (255, 223) Reed-Solomon decoder. The decoder is divided into twelve blocks as described below.

(1) *Syndrome Transform*: This block calculates the syndromes from the received 255 symbol messages. The output of this block is the syndrome polynomial $S(Z)$ (see Fig. 1). Figure 4 shows a more detailed diagram of the syndrome transform block. The architecture of this block is similar to that of the existing Reed-Solomon encoder chips (Ref. 3) since one of the multiplicands is fixed. Therefore the Berlekamp multiplier is used here. This finite field multiplier design is simple and a large number of gates are saved as compared to other schemes (Ref. 8). The calculated syndromes are fed to the Polynomial Expansion I block in parallel as the design of polynomial expansion circuit needs to load the multiplicand polynomial all at once. A parallel to serial conversion circuit is therefore saved. The number of transistors needed is about

$$\underset{\text{Registers}}{32 \times 48} + \underset{\text{Multipliers}}{32 \times 16} + \underset{\text{XOR's}}{32 \times 10} = 2{,}300$$

where we assume one symbol register contains 48 transistors, 16 transistors are required in a basic cell of a dual basis multiplier, and there are 32 XOR's with each XOR containing 10 transistors (Ref. 8). The previous syndrome computation circuit (Ref. 6) needed about

$$32 \times 400 = 12{,}800$$

transistors, since 32 syndromes are calculated and each syndrome cell contains 400 transistors.

(2) *Power Calculation*: This block converts the input erasure data into a sequence of α^k's and 0's. Since the maximum erasure decoding capability of this decoder is 32, only 32 symbol latches are needed here. Figure 3 shows a block diagram of the power calculation block. The multiplier used is a standard basis multiplier with a fixed multiplicand α. This is because output of this block must be in the standard basis and because the circuitry for Berlekamp's multiplier with the added dual-to-standard basis conversion is more complex than that simply using the slightly more complex standard basis multiplier only. A detection circuit for detecting the occurrence of erasures is included. If an erasure in the k_ith location occurs, its corresponding symbol α^{k_i} is calculated and latched. A counter is used to count the number of erasures. If this number exceeds 32, a decoding failure alarm will result and the received message will be passed without decoding. The number of transistors needed in this block is about

$$\underset{\text{Registers}}{32 \times 48} + \underset{\text{Power Calculation}}{500} = 2{,}000$$

where we assume one symbol register contains 48 transistors and 500 transistors are needed in the power calculation circuitry.

(3) *Polynomial Expansion I*: This block performs polynomial multiplication of the syndrome polynomial $S(Z)$ and the erasure locator polynomial $\Lambda(Z)$ to obtain the Forney syndrome polynomial (see Fig. 1). Figure 5 shows a detailed block diagram of the polynomial expansion block. In Fig. 5, the multiplicand polynomial is entered in parallel while the multiplier polynomial comes in bit by bit. The output Forney syndrome polynomial is fed to the modified Euclid's algorithm stage serially as required in Ref. 5. Therefore a parallel to serial conversion circuit is included. The number of transistors needed in this block is about

$$32 \times 350 = 11{,}200$$

This is because the maximum degree of the Forney syndrome polynomial is 31 for this code. Therefore 32 subcells for 32 coefficients are needed. We assume 350 transistors are used in each subcell.

(4) *Delay I*: This block delays the erasure locator polynomial $\Lambda(Z)$ to synchronize it with the error locator polynomial $\Lambda(Z)$ which comes out of the modified Euclid's algorithm unit (see Fig. 1). This block consists of a series of shift registers. An external signal is used to control the latch operation, i.e., $\Lambda(Z)$ is latched for a certain amount of time and then released to the next stage when the error locator polynomial is ready. Figure 6 presents the block diagram of this part. Since erasure locations come into the chip continuously, 5-stage multiplexing (see Ref. 7) is anticipated in this design. The number of transistors needed is then about

$$5 \times 32 \times 8 \times 8 = 10{,}240$$

where we assume five 32-symbol latches are needed. A symbol latch contains 8 bits and each bit has 8 transistors.

(5) *Polynomial Expansion II*: This block performs the polynomial multiplication of erasure locator polynomial

$\Lambda(Z)$ and error locator polynomial $\lambda(Z)$ to obtain the errata polynomial $P(Z)$. The errata polynomial is fed to the next stages in parallel (see Fig. 1). This block is similar to Polynomial Expansion I except that the parallel to serial conversion circuit is not needed here. Figure 7 shows its block diagram. The number of transistors needed is about

$$32 \times 300 = 9{,}600$$

where we assume 300 transistors are contained in each subcell and there are 32 subcells.

(6) *Modified Euclid's Algorithm*: This block performs the modified Euclid's algorithm. It was calculated in Ref. 7 that only 5 GCD (greatest common divisor) subcells are needed instead of the 32 subcells required in Ref. 6 (see also Appendix A). This is because a multiplexing method is used. Figure 8 shows the block diagram of a multiplexed GCD cell. Since each GCD subcell contains about 4000 transistors (Ref. 5), this block contains about $5 \times 4000 = 20{,}000$ transistors. The outputs are the error locator polynomial $\lambda(Z)$ and the errata evaluator polynomial $A(Z)$ (see Fig. 1). The error locator polynomial is fed to the Polynomial Expansion II in parallel, while the errata evaluator polynomial is fed to Polynomial Evaluation bit by bit.

(7) *Polynomial Evaluation*: This block performs the evaluation of the errata evaluator polynomial $A(Z)$. Figure 9 shows the block diagram of this circuit. The polynomial $A(Z)$ is the input of this block; outputs are the evaluated values of the $A(X_k)$'s. Because the architecture of this circuit is similar to that of syndrome calculation, the number of transistors needed in this block is about

$$2{,}300 + 2{,}400 = 4{,}700$$

where 2400 transistors are used for implementing the summation operation. The $A(X_k)$'s are fed into the next stage serially.

(8) *Derivative, Evaluation, Multiplication, and Inverse (DEMI)*: This block takes the derivative of the polynomial $P(Z)$, and performs the evaluation, multiplication by X_k, and inverse of the final product. Figure 10 shows its block diagram. The derivative of the polynomial $P(Z)$ is calculated by merely dropping the coefficients of even terms since the field in which the operations take place is of characteristic 2 (Ref. 7). Evaluation of the polynomial $P'(Z)$ is similar to the Polynomial Evaluation block except that there are only 16 coefficients in $P'(Z)$ – only half that in the Polynomial Evaluation block. Hence about 2,350 transistors are needed for this polynomial evaluation. In total, the number of transistors needed in this block is about

$$2{,}350 + 2{,}500 = 4{,}850$$

where we assume 2,500 transistors are needed for other parts. The output is fed to the next stage serially.

(9) *Chien Search*: This block performs the Chien search algorithm for both the error and erasure locations. The outputs are the error and erasure locations. This circuit is similar to that of the Polynomial Evaluation block since the Chien search algorithm actually is a polynomial evaluation process (Ref. 10). Hence the number of transistors needed in this block is about 4,700. The estimated error and erasure locations are fed to the next stage serially.

(10) *Delay II*: This block is the delay for received messages which are used together with estimated errors and erasures to obtain the estimated infomation. Because the received messages are fed into the chip serially and continuously, a pipeline register array is needed. This means that the delay cannot be multiplexed in the same way as the GCD operation. The number of transistors needed is about

$$1{,}343 \times 48 = 65{,}000$$

where 1,343 symbol delays are estimated to be needed and one symbol delay contains about 48 transistors. This block occupies almost half of the area of the decoder.

(11) *Delay III*: This block is used to delay the $A(X_k)$'s for synchronization with the output from the DEMI block. These two are sent to a multiplier to form the errata magnitude (see Fig. 1). It is estimated that 16 symbol delays are required; hence the number of transistors needed in this block is

$$16 \times 48 = 768$$

where 48 transistors are needed in one symbol delay.

(12) *Decoding Failure Detection*: This circuit performs the decoding failure detection. The algorithm for the decoding failure test will not be described here. A description of this algorithm may be found in Ref. 9. Suppose the Hamming weight of the errata vector computed by the decoder is w and that t errors and s erasures have occurred. The decoder has erred if $2w >$

$d + s$, where d is the design distance of this Reed-Solomon code, which is 33. It is estimated that this circuit contains about 1,500 transistors.

The total number of transistors needed to implement the (255, 223) Reed-Solomon decoder, if we exclude Delay II, is about

$$2{,}300 + 2{,}300 + 11{,}200 + 10{,}240 + 9{,}600$$

$$+ 20{,}000 + 4{,}700 + 4{,}850 + 4{,}700 + 768$$

$$+ 1{,}500 = 72{,}000$$

If overheads are included, the maximum number of transistors should not exceed 75,000. However, if Delay II is included in the calculation, 75,000 + 65,000 = 140,000 transistors are needed for this decoder. This would greatly reduce the probability of a single chip implementation using the fabrication processes that are currently open to JPL. However, the Delay II block, while being the largest single block in the design, is also the simplest and the most easily implemented with off-the-shelf technology. The whole block can be realized with a single medium density random access memory (RAM) chip.

IV. Improvements Over the Previous Design

Basically, the VLSI architecture of this new (255, 223) Reed-Solomon decoder is similar to that in Ref. 7. Figure 2 presents the VLSI architecture of the previous Reed-Solomon decoder. Several important improvements have been achieved which substantially increase the performance of the system. The improvements include the following:

(1) The long delay line in the input stage for storing the erasure locations is now eliminated (see Fig. 1). This results from the fact that the Chien search procedure searches for both error and erasure locations while the previous design searches only for error locations (Ref. 7; see Fig. 2). This design alleviates the necessity of storing erasure locations. About 1,200 symbol delays are reduced by this revision. When represented in terms of transistor numbers, it is approximately

$$1{,}200 \times 8 \times 6 = 57{,}600$$

where we assume one symbol register contains 8 bits and each bit has 6 transistors.

(2) Two polynomial multiplication blocks used in Ref. 7 are replaced by two polynomial expansion circuits (see Fig. 1). By this revision, the polynomial expansion block in the input stage of erasure locations in Ref. 7 is eliminated. This means a reduction of about 12,800 transistors.

(3) The power calculation block (this block calculates powers of symbols according to erasure locations) which was not presented in Ref. 7 is refined here. Figure 3 shows its block diagram. The 255 symbol latches needed in the previous design are now reduced to 32. This is due to the fact that an erasure detection circuit is added to detect the occurrence of erasures. If an erasure occurs, its location will be latched and moved one symbol forward if the next erasure occurs; otherwise it remains latched. Since the erasure correcting capability of this code is 32, only 32 symbol latches are needed. If the number of erasures occurred is greater than 32, a decoding failure alarm will be given. This saves 223 symbol latches, i.e., $223 \times 48 = 10{,}704$ transistors since each symbol latch contains 48 transistors.

(4) Berlekamp's multiplication algorithm (Ref. 8) is used in this decoder except in the power calculation block. It was discovered in Ref. 8 that the dual basis multiplication algorithm is the simplest among all known finite field multipliers. Also, the conversion between dual basis and standard basis is not complicated. Therefore if basis conversion is not used too often, the dual basis multiplier is the best choice. This is indeed the case in the Reed-Solomon decoder design.

Although the normal basis multiplier was used in the GCD design (Ref. 5), the revision is simple. The only modification is to replace the normal basis multiplier by the dual basis multiplier. The remainder of the circuit will be left unchanged. By this revision, basis conversions from normal to standard basis and dual basis are totally eliminated.

Due to the simplicity of Berlekamp's multiplication algorithm, the number of gates used in this design represents a substantial reduction over the previous ones. It is estimated that a Berlekamp general purpose multiplier needs about 400 transistors, while the Massey-Omura general purpose multiplier needs about 500 transistors. Therefore, if the multipliers are used frequently, the savings in the number of transistors is tremendous. In blocks such as Syndrome Transform, Polynomial Evaluation, and Chien Search where one of the multiplicands is fixed, the advantage of using the dual basis multiplier is more predominant.

Since there is only one multiplier in the power calculation circuit, the output must be in standard basis for compatibility

with next stage. For this reason, a standard basis multiplier is used in this block.

V. Implementation Procedure

The first step in the implementation of this decoder has already been accomplished. The various algorithms and architectures have been tested through analysis, simulation, and, in many cases, actual VLSI implementation. The next step is to implement versions of the various blocks so as to develop a complete Reed-Solomon decoding system. We plan to build the blocks described above on several separate chips. This will enhance their testability. When they are fully tested, the next step will be to integrate these working chips into 4 subchips as categorized in the following:

CHIP 1: This chip contains Delay II. The number of transistors needed is about 65,000. This chip may very well end up being an off-the-shelf RAM chip.

CHIP 2: This chip contains Syndrome Transform, Power Calculation, Polynomial Expansion I, and Delay I. The number of transistors is about

$$2{,}300 + 2{,}900 + 9{,}600 + 10{,}240 = 24{,}000$$

CHIP 3: This chip performs the modified Euclid's algorithm and contains about 20,000 transistors.

CHIP 4: This chip contains the following circuits: Polynomial Expansion II, Polynomial Evaluation, Chien Search, DEMI, Delay III, and Decoding Failure Detection. The number of transistors needed is about

$$9{,}600 + 4{,}400 + 4{,}400 + 4{,}700 + 768 + 1{,}500 = 25{,}400$$

When all of the above four chips are proven to work together as a system, the final step will be to put the whole decoder on a single silicon chip. It will be the first single VLSI chip (255, 223) Reed-Solomon decoder.

References

1. "Recommendation for Space Data System Standards: Telemetry Channel Coding," (Blue Book), Consulatative Committee for Space Data System, NASA, May 1984.
2. Miller, R. L., Deutsch, L. J., and Butman, S. A., "On the Error Statistics of Viterbi Decoding and the Performance of Concatenated Codes," *JPL Publication 81-3*, Jet Propulsion Laboratory, Pasadena, Calif., Sept. 1981.
3. Hsu, I. S., Reed, I. S, Truong, T. K., Wang, K., Yeh, C. S., and Deutsch, L. J., "The VLSI Implementation of a Reed-Solomon Encoder Using Berlekamp's Bit-Serial Multiplier Algorithm," *IEEE Trans. on Computers*, Vol. c-33, No. 10, pp. 906–911, Oct. 1984.
4. Brent, R. P., and Kung, H. T., "Systolic VLSI Arrays for Polynomial GCD Computations," Dept. Computer Science, Carnegie-Melon University, Pittsburgh, PA, Rep., 1982.
5. Hsu, I. S., Deutsch, L. J., Shao, H. M., and Truong, T. K., "A Single VLSI Chip for Polynomial GCD Computation," to be published in *TDA Progress Report*, Jet Propulsion Laboratory, Pasadena, Calif.
6. Shao, H. M., Truong, T. K., Deutsch, L. J., Yuen, J. H., and Reed, I. S., "A VLSI Design of a Pipeline Reed-Solomon Decoder," *IEEE Trans. on Computers*, Vol. c-34, No. 5, pp. 393–403, May 1985.
7. Shao, H. M., Truong, T. K., Hsu, I. S., Deutsch, L. J., and Reed, I. S., "A Single Chip VLSI Reed-Solomon Decoder," submitted to *IEEE Trans. on Computers*.

8. Hsu, I. S., Truong, T. K., Shao, H. M., Deutsch, L. J., and Reed, I. S., "A Comparison of VLSI Architecture of Finite Field Multiplier Using Dual, Normal, and Standard Basis," submitted to *IEEE Trans. on Computers*.

9. Miller, R. L., Truong, T. K., and Reed, I. S., "A Decoding Failure Test for the Transform Decoder of Reed-Solomon Codes," *TDA Progress Report 42-62*, Jet Propulsion Laboratory, Pasadena, Calif., pp. 121–124, Jan. 1981.

10. Peterson, W. W., and Weldon, E. L., *Error Correcting Codes*. Cambridge, Massachusetts: MIT press, 1980.

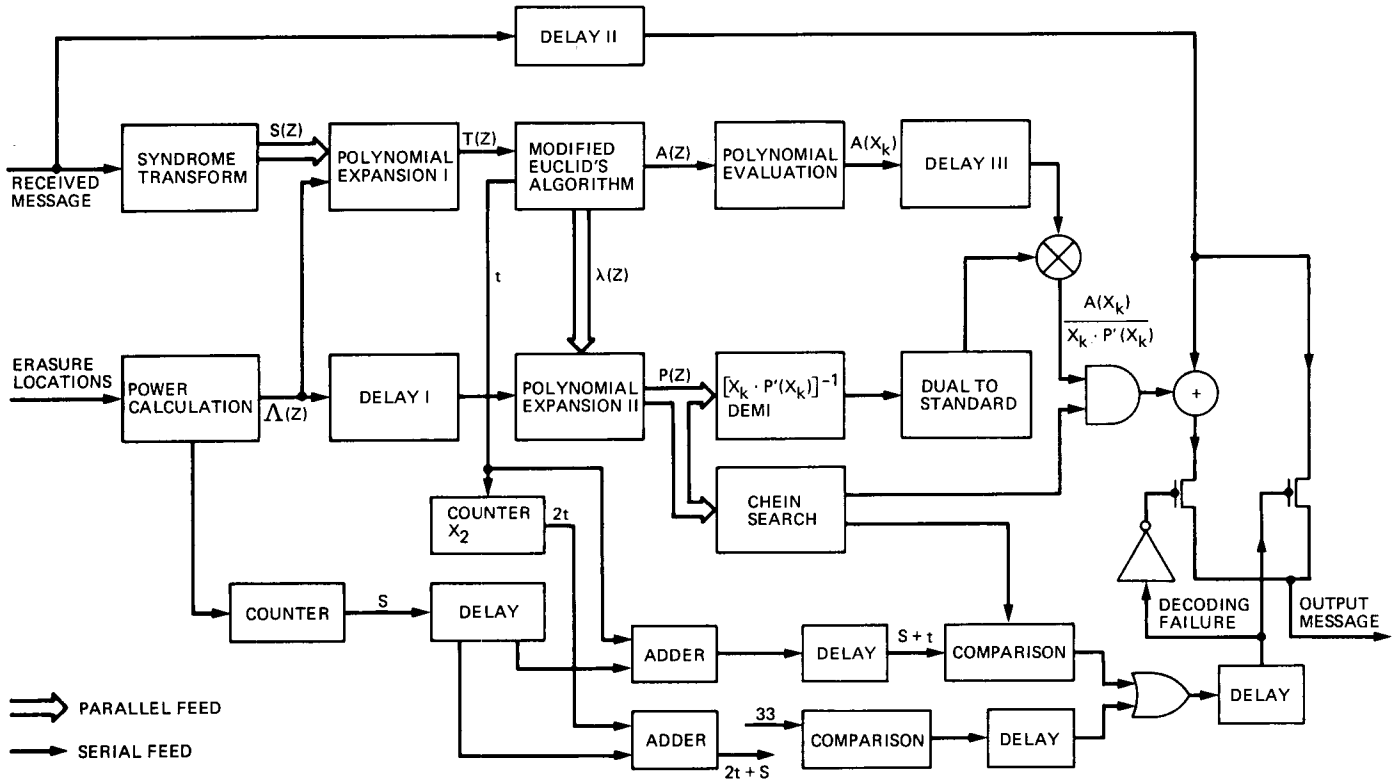


Fig. 1. Block diagram of the (255, 223) Reed-Solomon decoder for decoding both errors and erasures with decoding failure detection

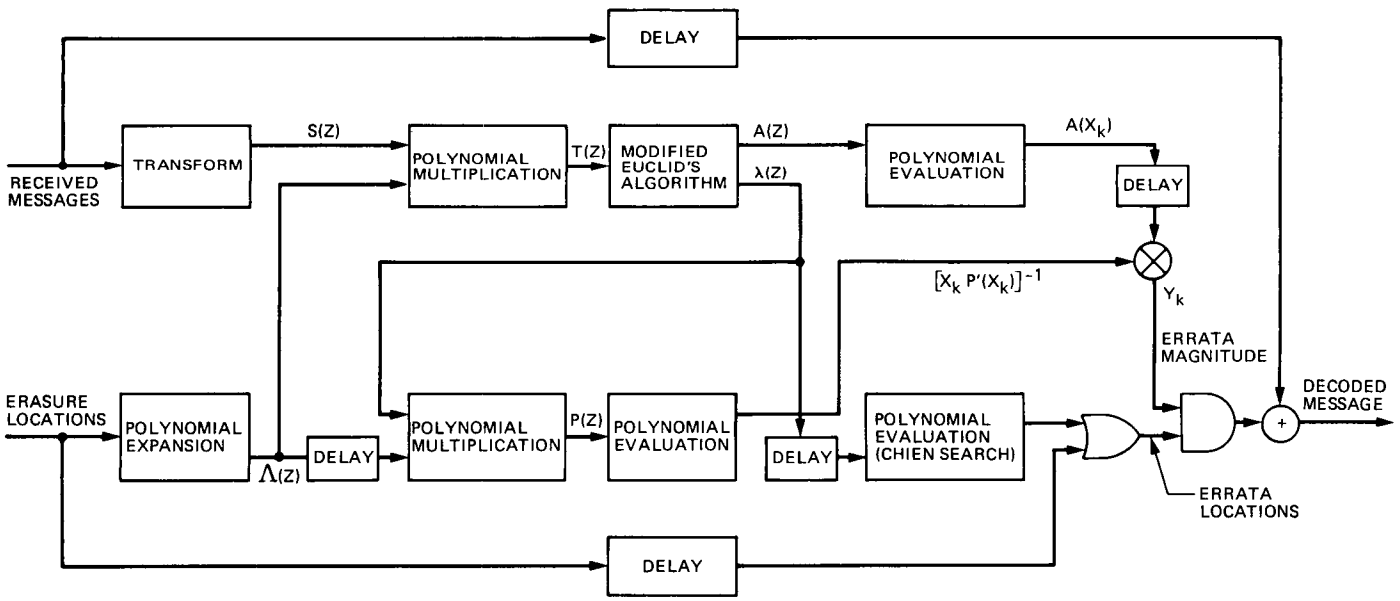


Fig. 2. VLSI architecture of the previously designed pipelined Reed-Solomon decoder for both error and erasure corrections

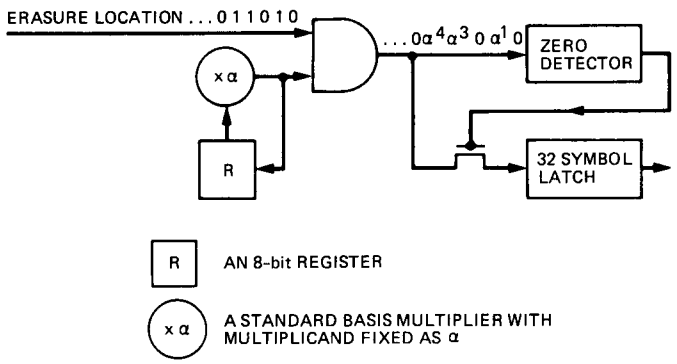


Fig. 3. The block diagram of the Power Calculation part

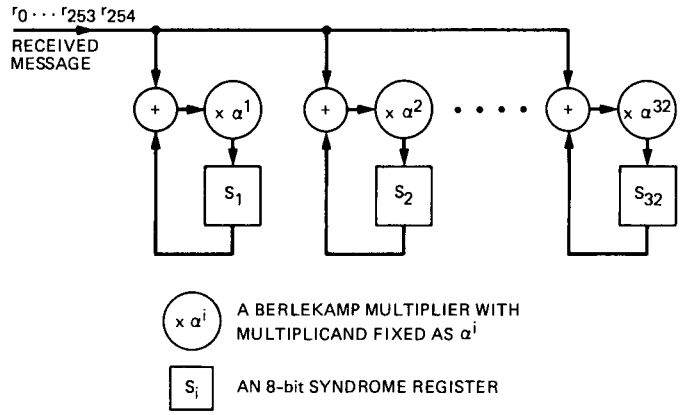


Fig. 4. The logic diagram of a Syndrome Transform block

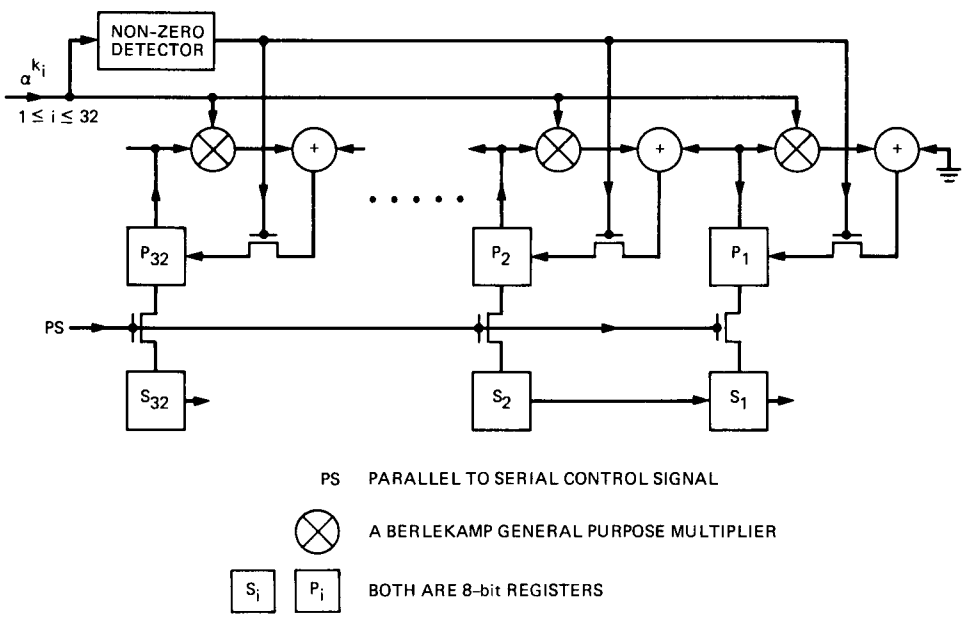


Fig. 5. The logic diagram of Polynomial Expansion I

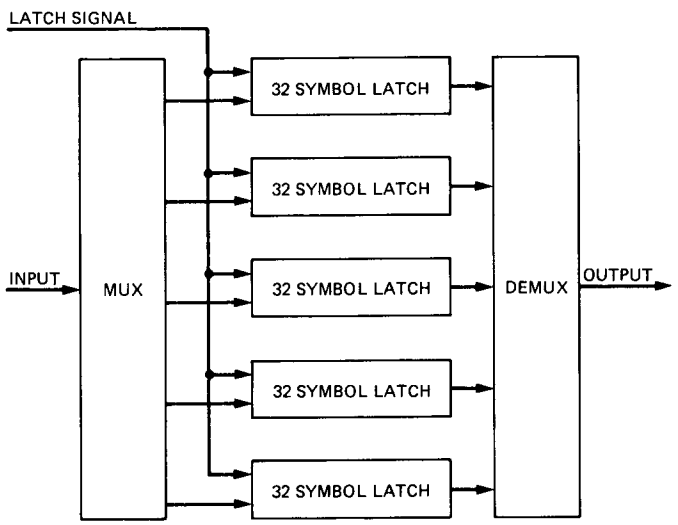


Fig. 6. The block diagram of Delay I

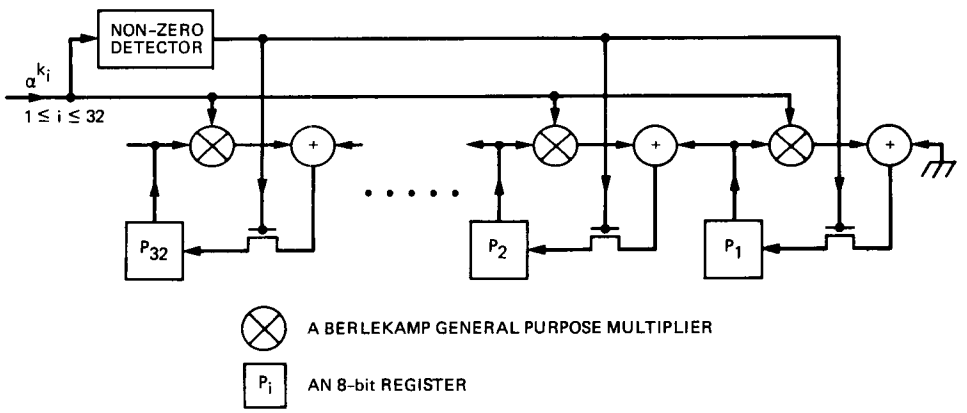


Fig. 7. The logic diagram of Polynomial Expansion II

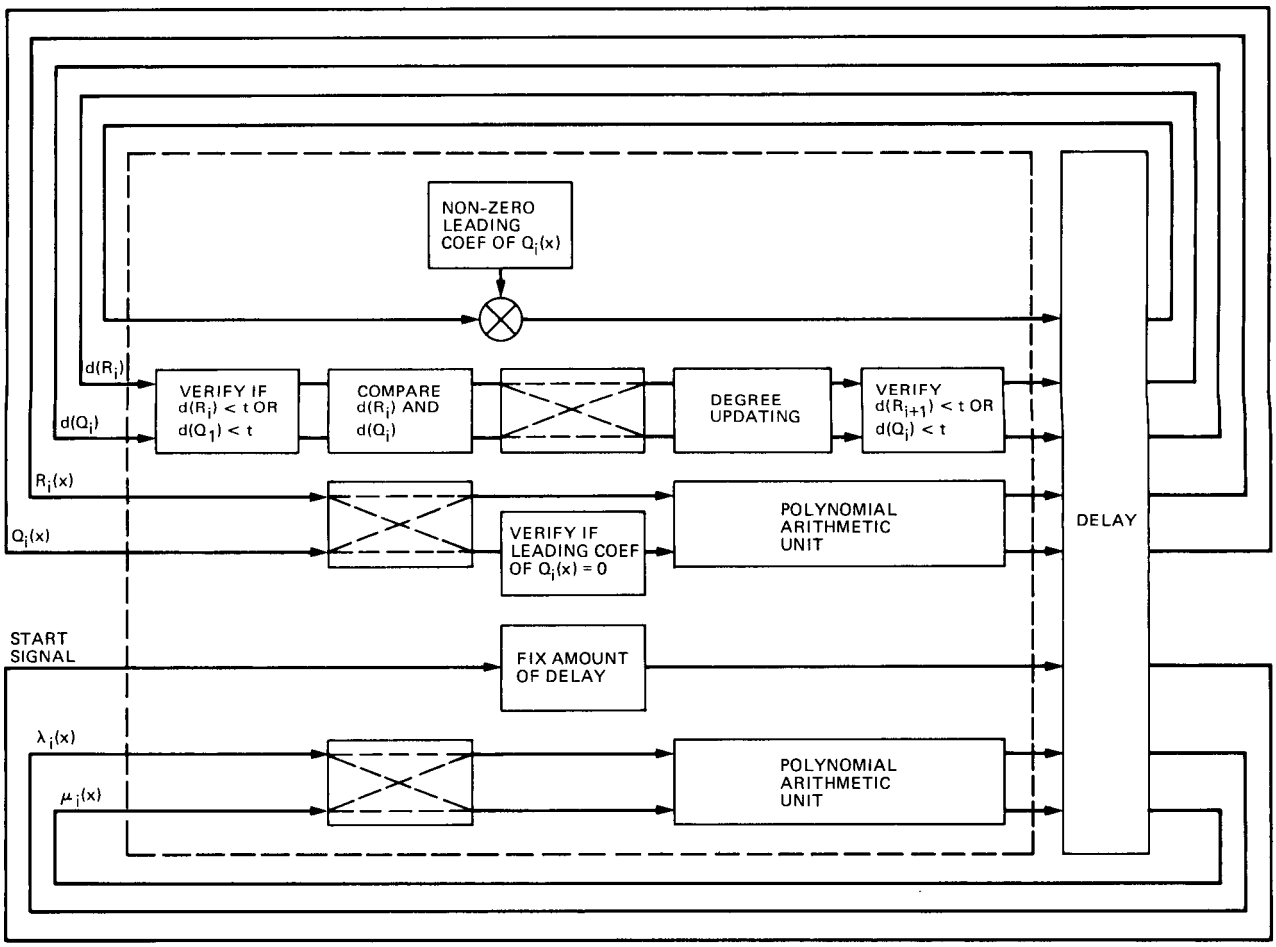


Fig. 8. The block diagram of a multiplexed GCD subcell

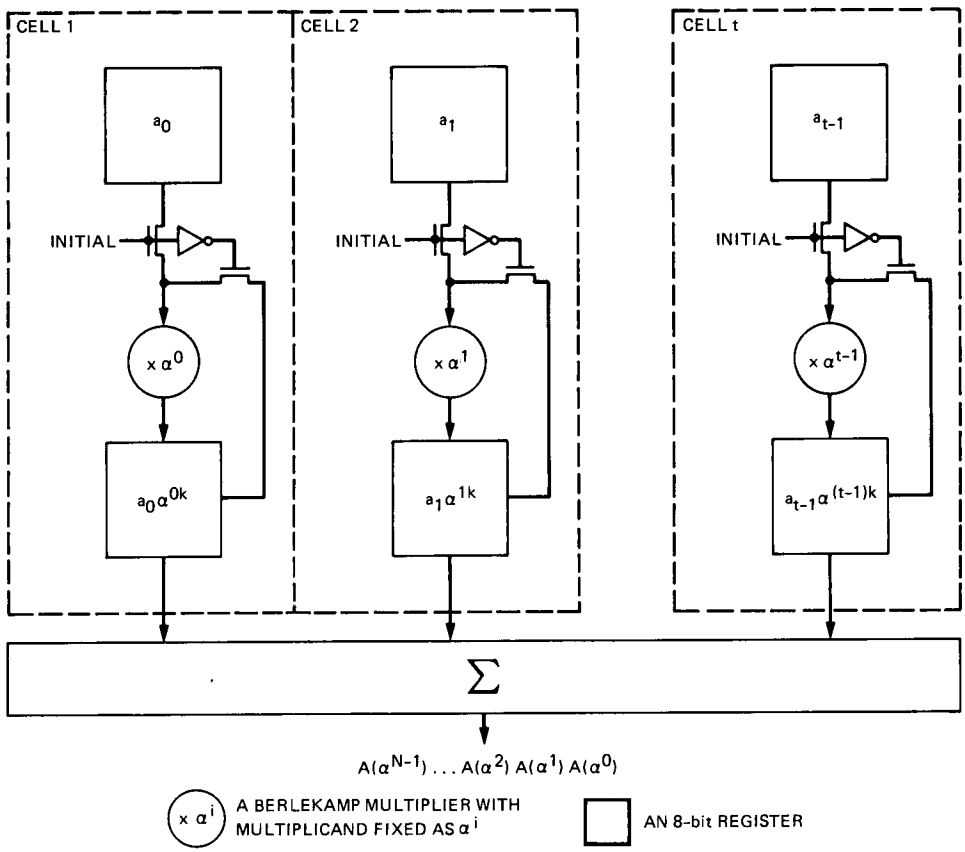


Fig. 9. The block diagram of the Polynomial Evaluation part

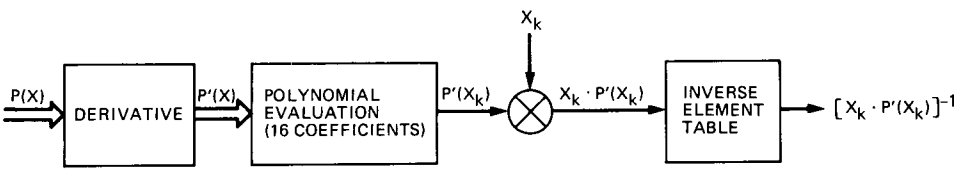


Fig. 10. The block diagram of DEMI

Appendix A

Comparison of the New Euclid's Algorithm With the Old

In this appendix, a comparison between the previous fully pipelined Reed-Solomon decoder design and the new multiplexed Reed-Solomon decoder design as described in this article is exhibited. The previous pipeline design of GCD cells needs about

$$\underset{\text{GCD Subcells}}{32 \times 4{,}000} + \underset{\text{Delays}}{160 \times 48} = 135{,}680$$

transistors, since 32 GCD subcells and 160 symbol delays are needed in the modified Euclid's algorithm block (Ref. 6). Figure A-1 shows the block diagram of the multiplexed modified Euclid's algorithm. For this new Euclid's algorithm, the number of transistors needed is about

$$\underset{\text{GCD Subcells}}{5 \times 4{,}000} + \underset{\text{Delays}}{5 \times 32 \times 8 \times 8} = 30{,}240$$

since only 5 stages of GCD subcells and five 32-symbol delays are needed. Apparently, the new multiplexed GCD design is much simpler than the previous pipelined GCD design.

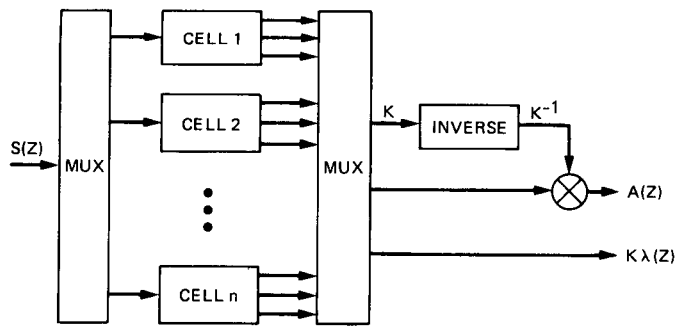


Fig. A-1. The block diagram of the multiplexed modified Euclid's algorithm

Appendix B

Transistor Count for the Transform Design

In this appendix, the number of transistors needed for using the transform algorithm to design a Reed-Solomon decoder is estimated. It will show that the transform decoding algorithm uses more transistors than the new algorithm as described in this article.

The VLSI architecture using transform methods to decode the Reed-Solomon code is quite different from that described in this article. Figure B-1 exhibits the VLSI architecture of such a Reed-Solomon decoder. Blocks such as Polynomial Evaluation, DEMI, Chien Search, and Delay III which are used in the present design are not needed. However, circuits for calculating Extended Syndrome from the coefficients of the combined error and erasure locator polynomial as well as syndrome, and inverse transform of error patterns are included. Also, the syndrome delay is necessary in this design for the reason stated above. It is estimated that this delay needs 6 stages of multiplexing. Hence the number of transistors needed for this part is about $6 \times 32 \times 8 \times 8 = 12{,}300$.

Since 32 cells are needed in the transformed error pattern calculation, and each cell needs about 400 transistors, $32 \times 400 = 12{,}800$ transistors are needed to implement this part. The architecture for the inverse transform of error patterns is similar to that of calculating syndrome except that 255 cells are needed instead of 32. That is to say, the inverse transform circuit is approximately 8 times larger than the syndrome calculation circuit. Hence the number of transistors needed for the inverse transform is about $8 \times 2{,}300 = 18{,}400$.

The total number of transistors needed for implementing the (255, 223) Reed-Solomon decoder using the transform decoder algorithm is about

$$2{,}300 + 2{,}300 + 9{,}600 + 10{,}240$$

$$+ 20{,}000 + 1{,}500 + 9{,}600 + 12{,}300$$

$$+ 12{,}800 + 18{,}400 = 98{,}440$$

If overhead is included, the number should not exceed 10,000 while the present design needs only 75,000 transistors. Note that the input message delay, Delay II, is not counted in this estimation.

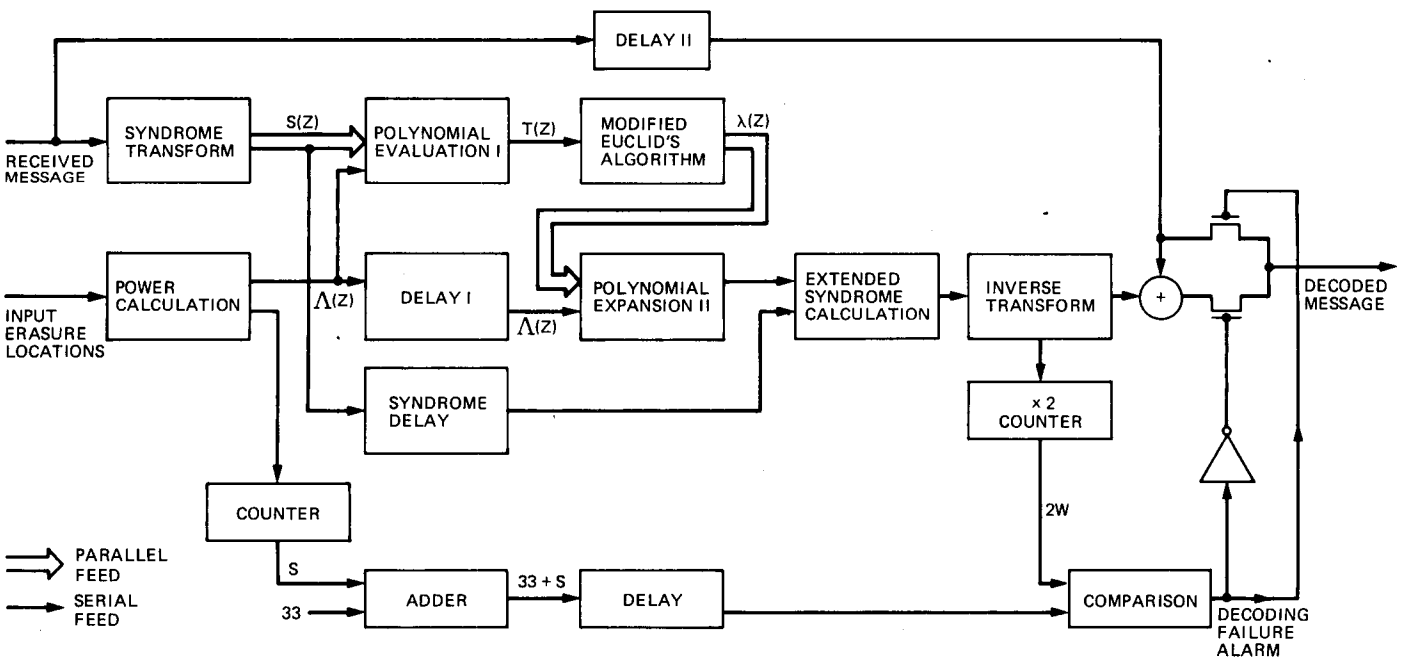


Fig. B-1. The VLSI architecture of a Reed-Solomon decoder using the transform decoding algorithm

A Simplified Procedure for Correcting Both Errors and Erasures of a Reed–Solomon Code Using the Euclidean Algorithm

T. K. Truong and I. S. Hsu
Communications Systems Research Section

W. L. Eastman
Mitre Corporation

I. S. Reed
University of Southern California

It is well known that the Euclidean algorithm or its equivalent, continued fractions, can be used to find the error locator polynomial and the error evaluator polynomial in Berlekamp's key equation needed to decode a Reed–Solomon (RS) code. In this article, a simplified procedure is developed and proved to correct erasures as well as errors by replacing the initial condition of the Euclidean algorithm by the erasure locator polynomial and the Forney syndrome polynomial. By this means, the errata locator polynomial and the errata evaluator polynomial can be obtained, simultaneously and simply, by the Euclidean algorithm only. With this improved technique the complexity of time-domain RS decoders for correcting both errors and erasures is reduced substantially from previous approaches. As a consequence, decoders for correcting both errors and erasures of RS codes can be made more modular, regular, simple, and naturally suitable for both VLSI and software implementation. An example illustrating this modified decoding procedure is given for a (15, 9) RS code.

I. Introduction

The Euclidean algorithm for solving the key equation for decoding both BCH and Goppa type codes was developed first by Sugiyama *et al.* [1]. Forney [2] defined an errata locator polynomial using what are now called Forney syndromes to correct both errors and erasures. Blahut [3, p. 258] showed that the errata locator polynomial can be computed directly by initializing Berlekamp's algorithm with the erasure locator polynomial.

Recently, Eastman[1] showed that the errata evaluator polynomial can be computed directly by initializing Berlekamp's

[1]W. L. Eastman, "Decoding Erasures," unpublished Mitre Corporation Report, Bedford, MA, 1986.

algorithm with the Forney syndrome polynomial. A proof of this new simplified decoding procedure is given in the present article. By this technique it is possible to compute the errata locator polynomial and the errata evaluator polynomial simultaneously from the Euclidean algorithm. It uses both the erasure locator polynomial and the Forney syndrome polynomial as initial conditions of the Euclidean algorithm.

This new Reed-Solomon (RS) decoder can be realized by a simplified pipeline architecture. An efficient time domain decoder can be developed for correcting both errors and erasures of RS codes. Such a decoding technique can be faster and simpler than previous methods [4].

II. The Time Domain Decoder for RS Codes

An algorithm is developed in [4] for time domain decoding RS codes to correct both errors and erasures by the use of continued fractions or its equivalent, the Euclidean algorithm. This algorithm is a modification of the Forney–Berlekamp method [2, 5]. The block diagram of such a decoding algorithm for a (255, 223) RS code over $GF(2^8)$ is depicted in Fig. 1. In this algorithm, the continued fraction algorithm is used to find the errata locator polynomial by replacing its initial condition by the erasure locator polynomial. The disadvantage of this algorithm is that after the errata locator polynomial $\tau(x)$ is obtained, by continued fractions, a polynomial multiplication is still needed to compute the errata evaluator polynomial $A(x) = \lfloor S(x)\,\tau(x) \rfloor$ from the known errata locator polynomial and the syndrome polynomial $S(x)$, where $\lfloor x \rfloor$ denotes the principal part of x.

In this section, the above-mentioned algorithm is modified to correct both errors and erasures in the time domain decoding of RS codes by a new use of the Euclidean algorithm. In this new algorithm, depicted in Fig. 2, the Euclidean algorithm is used to solve the Berlekamp–Forney key equation for the errata locator polynomial and the errata evaluator polynomial directly and simultaneously. The advantage of this algorithm over previous methods [4] is that the separate computation of the errata evaluator polynomial, usually needed as in [4], can be avoided. This new decoding algorithm is highly suitable to both VLSI and software implementation.

First, let $GF(2^m)$ be a finite field of 2^m elements. Also, let $N = 2^m - 1$ be the length of the (N, I) RS code over $GF(2^m)$ with minimum distance d, where $I = N - (d - 1)$ denotes the number of m-bit message symbols and $d - 1$ denotes the number of parity symbols such that $d - 1$ is either an even or an odd integer.

Define the following five vectors:

$c = (c_0, c_1, \ldots, c_{N-1})$, code vector

$r = (r_0, r_1, \ldots, r_{N-1})$, received vector

$e = (e_0, e_1, \ldots, e_{N-1})$, error vector

$u = (u_0, u_1, \ldots, u_{N-1})$, erasure vector

$u = (\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_{N-1})$, errata vector

These vectors are related by $\tilde{u} = e + u$ and $r = c + u + e$.

Suppose that t errors and v erasures occur in the received vector r and assume that $v + 2t \leqslant d - 1$. Next let α be a primitive element in $GF(2^m)$. Then $\gamma = \alpha^i$ is also a primitive element in $GF(2^m)$, where $(i, N) = 1$.

To minimize the complexity of an RS encoder, it is desirable that the generator polynomial be symmetric. If γ is a root of the code's generator polynomial, it is shown [6] that the generator polynomial $g(x)$ is symmetric, if and only if,

$$g(x) = \prod_{i=b}^{b+(d-2)} (x - \gamma^i) = \sum_{i=0}^{d-1} g_i x^i \qquad (1)$$

where $g_0 = g_{d-1} = 1$ and b satisfies the equality $2b + d - 2 = 2^m - 1$.

The syndromes of the code are given by

$$S_{(b-1)+k} = \sum_{i=0}^{N-1} \tilde{u}_i \gamma^{i(b-1+k)} = \sum_{j=1}^{v+i} Y_j X_j^{(b-1)+k}$$

$$\text{for } 1 \leqslant k \leqslant d-1 \qquad (2)$$

where X_j is either the j^{th} erasure or error location, and Y_j is either the j^{th} erasure or error magnitude. Define the sets, $\Lambda = \{X_i | X_i \text{ is an erasure location}\}$ and $\lambda = \{X_i | X_i \text{ is an error location}\}$. Also, it is not difficult to show, see [5], that

$$S(x) = \sum_{k=1}^{d-1} S_{(b-1)+k}\, x^{k-1}$$

$$= \sum_{j=1}^{v+t} \frac{Y_j X_j^b}{(1 - X_j x)} - \sum_{j=1}^{v+t} \frac{Y_j X_j^{b+d-1} x^{d-1}}{(1 - X_j x)} \qquad (3)$$

Following [4] define four different polynomials in terms of sets Λ and λ as follows:

The erasure locator:

$$\Lambda(x) = \prod_{X_j \epsilon \Lambda} (1 - X_j x) = \prod_{j=1}^{v} (1 - X_j x)$$

$$= \sum_{j=0}^{v} (-1)^j \Lambda_j x^j \tag{4a}$$

where $\Lambda_0 = 1$.

The error locator:

$$\lambda(x) = \prod_{X_j \epsilon \lambda} (1 - X_j x) = \prod_{j=1}^{t} (1 - X_j x)$$

$$= \sum_{j=0}^{t} (-1)^i \lambda_j x^j \tag{4b}$$

where $\lambda_0 = 1$.

The errata locator:

$$\tau(x) = \Lambda(x)\,\lambda(x) = \prod_{j=1}^{v+t} (1 - X_j x)$$

$$= \sum_{j=0}^{v+t} (-1)^j \tau_j x^j \tag{4c}$$

where $\tau_0 = 1$.

The errata evaluator:

$$A(x) = \sum_{j=1}^{v+t} Y_j X_j^b \left(\prod_{i \neq j} (1 - X_i x)\right) \tag{4d}$$

In terms of the polynomials defined above, (3) becomes

$$S(x) = \frac{A(x)}{\tau(x)} + \frac{x^{d-1} \sum_{j=1}^{v+t} Y_j X_j^{b+d-1} \left[\prod_{i \neq j} (1 - X_i x)\right]}{\tau(x)} \tag{5a}$$

or

$$S(x)\,\tau(x) = A(x) + x^{d-1} \sum_{j=1}^{v+t} Y_j X_j^{b+d-1} \left[\prod_{i \neq j} (1 - X_i x)\right] \tag{5b}$$

From (5b) one obtains the congruence relation

$$S(x)\,\tau(x) \equiv A(x) \bmod x^{d-1} \tag{6a}$$

Now define the set of formal power series

$$F = \left\{ \sum_{i=0}^{\infty} a_i x^{n-i} \mid a_i \in GF(2^m), \text{ and } n \text{ is an integer} \right\}$$

Since $\tau(x)$ in (6a) is a polynomial in x with the leading coefficient not equal to zero, it is not difficult to show that the inverse element $\tau^{-1}(x)$ always exists in the set of formal power series. Thus, (6a) can be solved for $S(x)$ to yield

$$S(x) \equiv \frac{A(x)}{\lambda(x)\,\Lambda(x)} \bmod x^{d-1} \tag{6b}$$

It is well known, e.g., see [5], that the maximum number of errors in an RS code which can be corrected is $\lfloor (d - 1 - v)/2 \rfloor$ where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x, i.e., the principal part of x. Now define a generalization of the Forney syndrome polynomial.

Definition 1. The Forney syndrome polynomial is defined by

$$T(x) \equiv S(x)\,\Lambda(x) \bmod x^{d-1} \tag{7}$$

By (7) and the key equation in (6b), $A(x)$ is

$$A(x) \equiv T(x)\,\lambda(x) \bmod x^{d-1} \tag{8a}$$

where

$$\deg\{\lambda(x)\} \leqslant \lfloor (d - 1 - v)/2 \rfloor$$

and

$$\deg\{A(x)\} \leqslant \lfloor (d + v - 3)/2 \rfloor \tag{8b}$$

Using a technique similar to that used for the proof of Theorem 7.7.3 in [3], one can prove an important theorem that the errata evaluator polynomial $A(x)$ and the errata locator polynomial $\tau(x)$ can be obtained simultaneously and

simply from the known $T(x)$, defined above, and the new key equation in (8). This algorithm takes into account both errors and erasures.

Theorem 7.7.1 in [3] can be used to solve for $A(x)$ and $\lambda(x)$ in (8). This theorem is the classical Euclidean algorithm for polynomials in matrix form. It is restated in terms of the polynomials and notation of the present article as follows:

Theorem 1: Given the two polynomials x^{d-1} and $T(x)$ in (7), where deg $\{x^{d-1}\} >$ deg $\{T(x)\}$. Let $M_0(x) = x^{d-1}$ and $R_0(x) = T(x)$. Also, let $N_s(x)$ be the 2×2 matrix equation which satisfies

$$N_s(x) = \begin{bmatrix} 0 & 1 \\ 1 & -Q_{s-1}(x) \end{bmatrix} N_{s-1}(x) \tag{9a}$$

where

$$Q_{s-1}(x) = \left\lfloor \frac{M_{s-1}(x)}{R_{s-1}(x)} \right\rfloor \tag{9b}$$

Finally, let

$$\begin{bmatrix} M_s(x) \\ R_s(x) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & -Q_{s-1}(x) \end{bmatrix} \begin{bmatrix} M_{s-1}(x) \\ R_{s-1}(x) \end{bmatrix} = N_s(x) \begin{bmatrix} x^{d-1} \\ T(x) \end{bmatrix} \tag{9c}$$

Then, for $s = r$, $M_s(x)$ satisfies

$$M_r(x) = c \,\text{GCD}\,[x^{d-1}, T(x)]$$

where c is a scalar, and $R_r(x) = 0$.

Proof: See [3, p. 194].

By Theorem 1 one observes that

$$N_s(x) = \prod_{t=0}^{s-1} \begin{bmatrix} 0 & 1 \\ 1 & -Q_t(x) \end{bmatrix} = \begin{bmatrix} A_s(x) & B_s(x) \\ C_s(x) & D_s(x) \end{bmatrix}. \tag{10a}$$

where

$$\deg\{D_s(x)\} > \deg\{B_s(x)\} \tag{10b}$$

It follows from (10a) that the determinant of $N_s(x)$ is $(-1)^s$. Thus, the inverse of $N_s(x)$ is given by

$$N_s^{-1}(x) = \begin{bmatrix} A_s(x) & B_s(x) \\ C_s(x) & D_s(x) \end{bmatrix}^{-1} = (-1)^s \begin{bmatrix} D_s(x) & -B_s(x) \\ -C_s(x) & A_s(x) \end{bmatrix} \tag{10c}$$

From (10a), $D_s(x)$ satisfies the recursive equation

$$D_s(x) = D_{s-2}(x) - Q_{s-1}(x) D_{s-1}(x) \tag{11}$$

for $s = 1, 2, \ldots$, where the initial conditions are $D_{-1}(x) = 0$ and $D_0(x) = 1$. Also from (9c), one obtains the equations

$$M_s(x) = R_{s-1}(x) \tag{12a}$$

$$R_s(x) = M_{s-1}(x) - Q_{s-1}(x) R_{s-1}(x) \tag{12b}$$

for $s = 1, 2, \ldots$, where

$$\deg\{M_s(x)\} > \deg\{R_s(x)\} \tag{12c}$$

Thus, from (12a) and (12b), the recursive formula for $R_s(x)$ is

$$R_s(x) = R_{s-2}(x) - Q_{s-1}(x) R_{s-1}(x) \tag{12d}$$

for $s = 1, 2, \ldots$, where the initial conditions are $R_{-1}(x) = x^{d-1}$ and $R_0(x) = T(x)$.

A substitution of (10a) into (9c) yields

$$\begin{bmatrix} M_s(x) \\ R_s(x) \end{bmatrix} = \begin{bmatrix} A_s(x) & B_s(x) \\ C_s(x) & D_s(x) \end{bmatrix} \begin{bmatrix} x^{d-1} \\ T(x) \end{bmatrix} \tag{13}$$

Thus, from (13), one obtains

$$R_s(x) \equiv T(x) D_s(x) \bmod x^{d-1} \tag{14}$$

for $s = 1, 2, \ldots$, where the recursive formulas for $D_s(x)$ and $R_s(x)$ are given in (11) and (12d), respectively.

Equation (14) is the form of Eq. (8a), which is used to solve for both $A(x)$ and $\lambda(x)$. To solve for $A(x)$ and $\lambda(x)$, one needs to find by (8b) a unique value $s = s'$, such that

$$\deg\{D_{s'}(x)\} \leq \lfloor (d - 1 - \nu)/2 \rfloor$$

and

$$\deg\{R_{s'}(x)\} \leqslant \lfloor((d-1-\nu)/2)-1\rfloor = \lfloor(\nu+d-3)/2\rfloor$$

in (14). Since $\deg\{R_{-1}(x)\} = d-1$ and $\deg\{R_s(x)\}$ is strictly decreasing with an increase in s, there always exists a unique value s' such that

$$\deg\{R_{s'-1}(x)\} \geqslant \left\lceil \frac{\nu+d-1}{2} \right\rceil \tag{15a}$$

and

$$\deg\{R_{s'}(x)\} \leqslant \left\lfloor \frac{\nu+d-3}{2} \right\rfloor \tag{15b}$$

where $\lceil x \rceil$ denotes the least integer greater than or equal to x.

Since $\deg\{D_{s'}(x)\}$ is increasing with an increase in s, then one needs to show also that

$$\deg\{D_{s'}(x)\} \leqslant \lfloor(d-1-\nu)/2\rfloor \tag{16}$$

To prove (16), it follows from (13) that

$$\begin{bmatrix} x^{d-1} \\ T(x) \end{bmatrix} = N_{s'}^{-1}(\lambda) \begin{bmatrix} M_{s'}(x) \\ R_{s'}(x) \end{bmatrix} \tag{17}$$

for $s = s'$, where $N_{s'}^{-1}(x)$ is the inverse of the 2×2 matrix $N_{s'}(x)$ in (10a). The substitution (10c) into (17) yields

$$\begin{bmatrix} x^{d-1} \\ T(x) \end{bmatrix} = (-1)^{s'} \begin{bmatrix} D_{s'}(x) & -B_{s'}(x) \\ -C_{s'}(x) & A_{s'}(x) \end{bmatrix} \begin{bmatrix} M_{s'}(x) \\ R_{s'}(x) \end{bmatrix} \tag{18}$$

From (18), one obtains

$$x^{d-1} = (-1)^{s'} [D_{s'}(x) M_{s'}(x) - B_{s'}(x) R_{s'}(x)] \tag{19a}$$

From (10b) and (12c), one knows that

$$\deg\{M_{s'}(x)\} > \deg\{R_{s'}(x)\}$$

and

$$\deg\{D_{s'}(x)\} > \deg\{B_{s'}(x)\}.$$

Thus, by (19a),

$$\deg\{x^{d-1}\} = \deg\{D_{s'}(x)\} + \deg\{M_{s'}(x)\} \tag{19b}$$

Since by (12a) for $s = s'$, $M_{s'}(x) = R_{s'-1}(x)$ and $\deg\{x^{d-1}\} = d-1$, then (19b) becomes

$$\deg\{D_{s'}(x)\} = d - 1 - \deg\{R_{s'-1}(x)\} \quad . \tag{20a}$$

By (15a), $\deg D_{s'}(x)$ in (20a) is upper bounded by the following inequality:

$$\deg\{D_{s'}(x)\} \leqslant d - 1 - \left\lceil \frac{\nu+d-1}{2} \right\rceil = \left\lfloor \frac{d-1-\nu}{2} \right\rfloor \tag{20b}$$

By (15b) and (20b), the Euclidean algorithm has the following stop conditions:

$$\deg\{R_{s'}(x)\} \leqslant \lfloor(\nu+d-3)/2\rfloor$$

and

$$\deg\{D_{s'}(x)\} \leqslant \lfloor(d-1-\nu)/2\rfloor$$

Thus, also by (15a) and (20b), polynomials $R_{s'}(x)$ and $D_{s'}(x)$ are solutions of (14). Uniqueness follows from the fact that there is only one solution to (8a) under the conditions of (8b). Therefore, $A(x) = R_{s'}(x)/\Delta$ and $\lambda(x) = D_{s'}(x)/\Delta$ are the unique solutions of (8a, b), where Δ is chosen to ensure that $\lambda_0 = 1$.

To obtain $\tau(x)$, one replaces the recursive equation in (11) by

$$\tau_s(x) = -Q_{s-1}(x)\,\tau_{s-1}(x) + \tau_{s-2}(x) \tag{21}$$

with the changed initial conditions: $\tau_0(x) = \Lambda(x)$ and $\tau_{-1}(x) = 0$. To obtain the final errata locator polynomial, first, let $s = 1$. Then (21) becomes by (11),

$$\tau_1(x) = -Q_0(x)\,\tau_0(x) + \tau_{-1}(x)$$

$$= -Q_0(x)\,\Lambda(x) = D_1(x)\,\Lambda(x)$$

For $s = 2$, (21) becomes by (11),

$$\tau_2(x) = -Q_1(x)\,\tau_1(x) + \tau_0(x) = -Q_1(x)\,D_1(x)\,\Lambda(x) + \Lambda(x)$$

$$= (-Q_1(x)\,D_1(x) + 1)\,\Lambda(x) = D_2(x)\,\Lambda(x)$$

etc. Here, in general, one has $\tau_s(x) = D_s(x)\,\Lambda(x)$ with $\tau_0(x) = \Lambda(x)$ and $\tau_{-1}(x) = 0$.

The results proved above imply the following now evident theorem:

Theorem 2: Let $T(x)$ in (7) be the Forney syndrome polynomial of a t-error and ν-erasure correcting RS code under the condition $\nu + 2t \leqslant d - 1$, where $d - 1$ is either an even or an odd integer. Consider the two polynomials x^{d-1} and $T(x)$ in (7). The Euclidean algorithm for polynomials on $GF(2^m)$ can be used to develop two finite sequences $R_s(x)$ and $\tau_s(x)$ from the following two recursive formulas:

$$\tau_s(x) = (-Q_{s-1}(x))\,\tau_{s-1}(x) + \tau_{s-2}(x) \tag{22a}$$

and

$$R_s(x) = R_{s-2}(x) - Q_{s-1}(x)\,R_{s-1}(x) \tag{22b}$$

for $s = 1, 2, \ldots$, where the initial conditions are $\tau_0(x) = \Lambda(x)$, $\tau_{-1}(x) = 0$, $R_{-1}(x) = x^{d-1}$, and $R_0(x) = T(x)$. Here $Q_{s-1}(x)$ is obtained as the principal part of $R_{s-2}(x)/R_{s-1}(x)$. The recursion in (22a) and (22b) for $R_s(t)$ and $\tau_s(x)$ terminates when $\deg\{R_s(x)\} \leqslant \lfloor (d + \nu - 3)/2 \rfloor$ for the first time for some value $s = s'$. Let

$$A(x) = R_{s'}(x)/\Delta \tag{23a}$$

and

$$\tau(x) = \tau_{s'}(x)/\Delta \tag{23b}$$

Also in (23) $\Delta = \tau_{s'}(0)$ is a field element in $GF(2^m)$ which is chosen so that $\tau_0 = 1$. Then $A(x)$ and $\tau(x)$ in (23) are the unique solution of $A(x) \equiv T(x)\,\tau(x) \bmod x^{d-1}$, where both inequalities, $\deg\{\tau(x)\} \leqslant \lfloor (d + \nu - 1)/2 \rfloor$ and $\deg\{A(x)\} \leqslant \lfloor (d + \nu - 3)/2 \rfloor$, are satisfied.

The proof of Theorem 2 demonstrates that the algorithm presented by Eastman[1] is correct.

The roots of $\tau(x)$ are the inverse locations of the t errors and ν erasures. These roots are most efficiently found by the Chien search procedure. By (4d) it is shown readily that the errata values are

$$Y_k = \frac{A(X_k^{-1})}{(X_k^{b-1}\,\tau'(X_k^{-1}))} \qquad \text{for } 1 \leqslant k \leqslant \nu + t \tag{24}$$

where $\tau'(X_k^{-1})$ is the derivative with respect to x of $\tau(x)$, evaluated at $x = X_k^{-1}$.

The overall time-domain decoding of RS codes for correcting errors and erasures, using Theorem 2 and the Euclidean algorithm, is summarized in the following steps:

(1) Compute the transform of the received vector m-tuple over $GF(2^m)$ from Eq. (2). Next calculate the erasure locator polynomial $\Lambda(x)$ from Eq. (4a) and define $\deg\{\Lambda(x)\} = \nu$.

(2) Compute the Forney syndrome polynomial from $T(x)$ in (7).

(3) To determine the errata locator polynomial $\tau(x)$ and errata evaluator polynomial $A(x)$, where $0 \leqslant \nu < d - 1$, apply the Euclidean algorithm to x^{d-1} and $T(x)$ as given by (7). The initial values of the Euclidean algorithm are $\tau_0(x) = \Lambda(x)$, $\tau_{-1}(x) = 0$, $R_{-1}(x) = x^{d-1}$ and $R_0(x) = T(x)$. For $\nu = d - 1$ set $\tau(x) = \Lambda(x)$ and $A(x) = T(x)$.

(4) Compute the errata values from (24).

To illustrate the time domain decoding procedure for correcting errors and erasures, an elementary example of an RS code over $GF(2^4)$ is now presented. The representation of the field $GF(2^4)$ generated by the primitive irreducible polynomial $G(x) = x^4 + x + 1$ is given in Table A-1 in the appendix.

Example 1: Consider a (15, 9) RS code over $GF(2^4)$ with minimum distance $d = 7$. In this code, ν erasures and t errors under the condition $2t + \nu \leqslant d - 1$ can be corrected. In order to simplify this example, let $\gamma = \alpha$ and $b = 1$. Thus, the generator polynomial of such a (15, 9) RS code is defined by

$$g(x) = \prod_{i=1}^{6}(x - \alpha^i) = x^6 + \alpha^{10}x^5 + \alpha^{14}x^4 + \alpha^4x^3 + \alpha^6x^2 + \alpha^9x + \alpha^6$$

Assume the message symbols are

$$I(x) = \alpha^{10}x^{14} + \alpha^{12}x^{13} + \alpha^8x^{12} + \alpha^5x^{11} + \alpha^6x^{10} + \alpha^{14}x^9 + \alpha^{13}x^8 + \alpha^{11}x^7 + \alpha^9x^6$$

The encoded codeword, which is a multiple of $g(x)$, is

$$c(x) = \alpha^{10}x^{14} + \alpha^{12}x^{13} + \alpha^8x^{12} + \alpha^5x^{11} + \alpha^6x^{10} + \alpha^{14}x^9 + \alpha^{13}x^8 + \alpha^{11}x^7 + \alpha^9x^6 + x^5 + \alpha x^4 + \alpha^2x^3 + \alpha^6x^2 + \alpha^{12}x + \alpha^8$$

Written as a vector, the codeword is

$$c = (\alpha^{10}, \alpha^{12}, \alpha^{8}, \alpha^{5}, \alpha^{6}, \alpha^{14}, \alpha^{13}, \alpha^{11}, \alpha^{9}, \alpha^{0}, \alpha, \alpha^{2}, \alpha^{6}, \alpha^{12}, \alpha^{8})$$

Assume the erasure vector is

$$u = (0, 0, 0, 0, 0, 0, 0, \alpha^{2}, 0, 0, 0, 0, 0, 0, 0)$$

and the error vector is

$$e = (0, 0, 0, 0, \alpha^{11}, 0, 0, 0, 0, 0, 0, \alpha^{7}, 0, 0, 0)$$

Then the errata vector is

$$\tilde{u} = u + e = (0, 0, 0, 0, \alpha^{11}, 0, 0, \alpha^{2}, 0, 0, 0, \alpha^{7}, 0, 0, 0)$$

Assume the received vector is

$$r = c + \tilde{u} = (\alpha^{10}, \alpha^{12}, \alpha^{8}, \alpha^{5}, \alpha, \alpha^{14}, \alpha^{13}, \alpha^{9}, \alpha^{9}, \alpha^{0}, \alpha, \alpha^{12}, \alpha^{6}, \alpha^{12}, \alpha^{8})$$

The syndromes S_k for r are

$$S_k = \sum_{n=0}^{14} r_n \alpha^{n \cdot k} = \alpha^{7}(\alpha^{3})^{k} + \alpha^{2}(\alpha^{7})^{k} + \alpha^{11}(\alpha^{10})^{k} \quad \text{for } 1 \leqslant k \leqslant 6$$

This yields $S_1 = \alpha^0$, $S_2 = \alpha^{13}$, $S_3 = \alpha^{14}$, $S_4 = \alpha^{11}$, $S_5 = \alpha$, and $S_6 = 0$. Thus, the syndrome polynomial is

$$S(x) = \alpha^{0} + \alpha^{13}x + \alpha^{14}x^{2} + \alpha^{11}x^{3} + \alpha x^{4} + 0x^{5}$$

The erasure locator polynomial is $\Lambda(x) = (1 + \alpha^{7}x)$. In this example, the maximum erasure correcting capability is

$$\lfloor (d - 1 - \nu)/2 \rfloor = \lfloor (7 - 1 - 1)/2 \rfloor = 2$$

By (7), one obtains the Forney syndrome polynomial as

$$T(x) \equiv \Lambda(x)S(x) \equiv (1 + \alpha^{7}x)(1 + \alpha^{13}x + \alpha^{14}x^{2} + \alpha^{11}x^{3} + \alpha x^{4} + 0x^{5}) \bmod x^{6}$$

$$= \alpha^{8}x^{5} + \alpha^{9}x^{4} + \alpha x^{3} + \alpha^{12}x^{2} + \alpha^{5}x + \alpha^{0} \tag{25}$$

In (25), the coefficients of $T(x)$, $T_0 = \alpha^0$, $T_1 = \alpha^5$, $T_2 = \alpha^{12}$, $T_3 = \alpha$, $T_4 = \alpha^9$, and $T_5 = \alpha^8$ are the Forney syndromes.

The Euclidean algorithm is applied next to polynomials x^{d-1} and $T(x)$ in (7). By this means, polynomials $\tau(x)$ and $A(x)$ are determined by use of the Euclidean algorithm. This is accomplished by the recursive formulas (22a) and (22b) illustrated in Table 1, where initially $R_{-1}(x) = x^{d-1} = x^6$ and $R_0(x) = T(x) = \alpha^{8}x^{5} + \alpha^{9}x^{4} + \alpha x^{3} + \alpha^{12}x^{2} + \alpha^{5}x + 1$. From Table 1, one observes that deg $\{R_{s'}(x)\}$ = deg $\{R_2(x)\} = 2 \leqslant \lfloor (d + \nu - 3)/2 \rfloor = 2$. Thus, the computation terminates at this point for $s' = 2$, and

$$R_2(x) = \alpha^{7}x^{2} + \alpha x + \alpha^{2}$$

and

$$\tau_2(x) = \alpha^{7}x^{3} + \alpha^{13}x^{2} + \alpha^{4}x + \alpha^{2}$$

By (23a) and (23b), one has

$$\tau(x) = \frac{1}{\alpha^{2}} \tau_2(x) = \alpha^{5}x^{3} + \alpha^{11}x^{2} + \alpha^{2}x + 1 \tag{26}$$

and

$$A(x) = \frac{1}{\alpha^{2}} R_2(x) = \alpha^{5}x^{2} + \alpha^{14}x + 1 \tag{27}$$

By use of a Chien search, the roots of $\tau(x)$ constitute the set $\{\alpha^{-7}, \alpha^{-3}, \alpha^{-10}\}$. The derivative with respect to x of $\tau(x)$ in (26) is $\tau'(x) = \alpha^{5}x^{2} + \alpha^{2}$. Thus, by (24) and (27), the errata values are

$$Y_1 = \frac{A(X_1^{-1})}{\tau'(X_1^{-1})} = \frac{A(\alpha^{-7})}{\tau'(\alpha^{-7})} = \frac{\alpha^{5}(\alpha^{-7})^{2} + \alpha^{14}(\alpha^{-7}) + 1}{\alpha^{5}(\alpha^{-7})^{2} + \alpha^{2}} = \alpha^{2}$$

$$Y_2 = \frac{A(X_2^{-1})}{\tau'(X_2^{-1})} = \frac{\alpha^{5}(\alpha^{-3}) + \alpha^{14}(\alpha^{-3}) + 1}{\alpha^{5}(\alpha^{-3})^{2} + \alpha^{2}} = \alpha^{7}$$

and

$$Y_3 = \frac{A(X_3^{-1})}{\tau'(X_3^{-1})} = \frac{\alpha^{5}(\alpha^{-10})^{2} + \alpha^{14}(\alpha^{-10}) + 1}{\alpha^{5}(\alpha^{-10})^{2} + \alpha^{2}} = \alpha^{11}$$

III. An Improved Architecture for the Time Domain Decoder

An improved VLSI architecture of a pipeline decoder for correcting both errors and erasures of a (255, 223) RS decoder for codes over $GF(2^8)$ is presented in Fig. 2. For this example, assume $\gamma = \alpha$ and $b = 112$.

In Fig. 2, the block diagram can be separated into two parts as indicated by the dashed line. The functional units contained in the first part of the decoder architectures in Fig. 2 are shown as follows: (1) the syndrome computation unit, (2) the power calculation unit, (3) the power expansion unit, (4) the polynomial expansion unit, (5) the $\lfloor (d + \nu - 3)/2 \rfloor$ generator, and (6) the delay unit.

The syndrome computation unit accepts received messages and computes their syndromes. There are 32 syndrome subcells in a (255, 223) RS decoder. The computed syndrome polynomial is labelled as $S(x)$ in Fig. 2. The coefficients of $S(x)$ are fed in parallel to the polynomial expansion unit in order to compute the Forney syndromes.

The power calculation unit converts the received 1's and 0's into a sequence of α^k's and 0's, where α is a primitive element of the finite field over which the RS code is defined. These received 1's and 0's indicate the occurrence or nonoccurrence, respectively, of an erasure at a specific location. Since the maximum erasure correcting capability of a (255, 223) RS decoder is 32, only 32 symbol latches are needed to store all correctable erasure locations.

A detection circuit for detecting the occurrence of erasures is included in the power calculation unit. If an erasure occurs at the kth location, a symbol α^k is calculated by the power calculation unit and latched. The sequence of α^k's is fed to the polynomial expansion circuit, to the power expansion unit and to the $\lfloor (d + \nu - 3)/2 \rfloor$ generator.

The power expansion unit converts the α^k's into an erasure locator polynomial $\Lambda(x)$. Therefore, the polynomial $\Lambda(x)$ has α^k's as its roots. The erasure locator polynomial $\Lambda(x)$ is fed to the modified GCD unit as one of the initial conditions for the modified GCD unit.

A generator is used to compute $\lfloor (d + \nu - 3)/2 \rfloor$. The output is sent to the modified GCD unit and used as a stop indicator for the Euclidean algorithm. The polynomial expansion unit is used to compute the required Forney syndromes.

In Fig. 2, the erasure locator polynomial $\Lambda(x)$, together with the Forney syndrome polynomial $T(x)$, is the input to the modified GCD unit. The outputs of the modified GCD unit are the errata locator polynomial, $\tau(x)$, and the errata evaluator polynomial, $A(x)$. The error correcting capability of the code is computed by $\lfloor (32 - \nu)/2 \rfloor$.

The second half of Fig. 2 is labelled as "II." One of the outputs of the modified GCD unit is the errata locator polynomial $\tau(x)$. This output is fed to a Chien search unit and to another unit for computing $[x^{b-1}\tau'(x)]^{-1} = [x^{111}\tau'(x)]^{-1}$, where $b = 112$. The other output of the modified GCD is the errata evaluator polynomial $A(x)$. This is fed to the polynomial evaluation unit to perform the evaluation of $A(x)$.

The $[x^{111}\tau'(x)]^{-1}$ unit computes one part of the errata magnitude. The product of the outputs from the polynomial evaluation unit and the $[x^{111}\tau'(x)]^{-1}$ unit forms the errata magnitude.

In Fig. 2, the Chien search unit is used to search for both the error and erasure locations. The architecture of the Chien search unit is similar to that of a polynomial evaluation unit, except that there is a zero detector at the end in the Chien search unit.

Compared with the previous design in Fig. 1, one observes that this improved architecture does not require the polynomial multiplication unit, delay II, delay III, and the truncation circuit usually needed in Fig. 1 for computing the errata evaluator polynomial. Thus, this new decoding algorithm in Fig. 2 is simpler and more suitable for the VLSI implementation. Finally, a comparison of VLSI architecture for time and transform domain decoding of Reed-Solomon codes is shown in [7].

References

[1] Y. Sugiyama, M. Kasahara, S. Hirasawa, and T. Namekawa, "A method for solving key equation for decoding Goppa codes," *Inf. and Contr.*, Vol. 27, pp. 87–99, 1975.

[2] G. D. Forney, "On decoding BCH codes," *IEEE Trans. on Information Theory*, Vol. IT-11, pp. 549–557, 1965.

[3] R. E. Blahut, *Theory and Practice of Error Control Codes*, Addison-Wesley Publishing Co., CA, p. 258, May 1984.

[4] I. S. Reed, T. K. Truong, and R. L. Miller, "Decoding of B.C.H. and RS codes with errors and erasures using continued fractions," *Electronics Letters*, Vol. 15, No. 17, pp. 542–544, August 16, 1976.

[5] E. P. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, 1968.

[6] E. R. Berlekamp, "Bit-Serial Reed-Solomon Encoders," *IEEE Trans. on Information Theory*, Vol. IT-28, No. 6, pp. 869–874, November 1982.

[7] T. K. Truong, I. S. Hsu, I. S. Reed, E. Satorius and L. J. Deutsch, "A Comparison of VLSI architecture for time and transform domain decoding of Reed-Solomon codes," presented at ICCD '87 Conference, New York, October 5-8, 1987.

Table 1. An example of the Euclidean algorithm used to find $\tau(x)$ and $A(x)$

S	$R_{s-2}(x) = Q_{s-1}(x)R_{s-1}(x) + R_s(x)$	$Q_{s-1}(x)$	$R_s(x)$	$\tau_s(x)$
-1			x^6	0
0			$\alpha^8x^5 + \alpha^9x^4 + \alpha x^3 + \alpha^{12}x^2 + \alpha^5x + 1$	$1 + \alpha^7x$
1	$\alpha^8x^5 + \alpha^9x^4 + \alpha x^3 + \alpha^{12}x^2 + \alpha^5x + 1 \overline{)\, x^6}$ $\frac{1}{\alpha^8}x + \frac{\alpha}{\alpha^8}$ $x^6 + \alpha x^5 + \alpha^8x^4 + \alpha^4x^3 + \alpha^{12}x^2 + \alpha^7x$ $\alpha x^5 + \alpha^8x^4 + \alpha^4x^3 + \alpha^{12}x^2 + \alpha^7x$ $\alpha x^5 + \alpha^2x^4 + \alpha^9x^3 + \alpha^5x^2 + \alpha^{13}x + \alpha^8$ $x^4 + \alpha^{14}x^3 + \alpha^{14}x^2 + \alpha^3x + \alpha^8$	$\frac{1}{\alpha^8}(x + \alpha)$	$x^4 + \alpha^{14}x^3 + \alpha^6x^2 + \alpha^2x + \alpha^8$	$\frac{1}{\alpha^8}(x + \alpha) \cdot (1 + \alpha^7x) + 0 = \frac{1}{\alpha^8}(\alpha^7x^2 + \alpha^2x + \alpha)$
2	$x^4 + \alpha^{14}x^3 + \alpha^{14}x^2 + \alpha^5x + \alpha^8 \overline{)\, \alpha^8x^5 + \alpha^9x^4 + \alpha x^3 + \alpha^{12}x^2 + \alpha^5x + 1}$ $\alpha^8x + 1$ $\alpha^8x^5 + \alpha^7x^4 + \alpha^7x^3 + \alpha^{13}x^2 + \alpha x$ $x^4 + \alpha^{14}x^3 + \alpha x^2 + \alpha^2x + 1$ $x^4 + \alpha^{14}x^3 + \alpha^{14}x^2 + \alpha^5x + \alpha^8$ $\alpha^7x^2 + \alpha x + \alpha^2$	$\alpha^8x + 1$	$\alpha^7x^2 + \alpha x + \alpha^2$	$(\alpha^7x^2 + \alpha^2x + \alpha) \cdot (x + \alpha^7) + (1 + \alpha^7x) = \alpha^7x^3 + \alpha^{13}x^2 + \alpha^4x + \alpha^2$

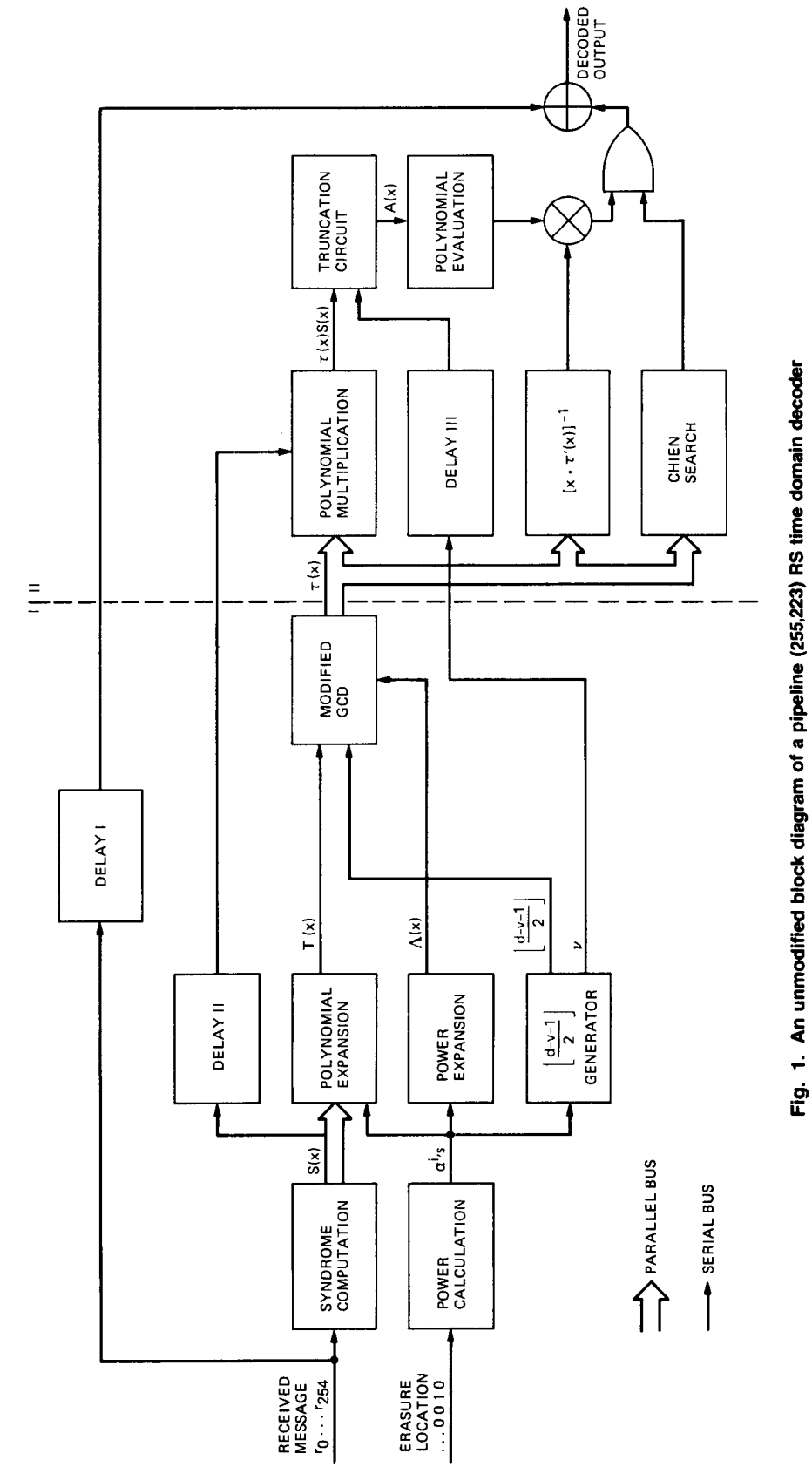


Fig. 1. An unmodified block diagram of a pipeline (255,223) RS time domain decoder

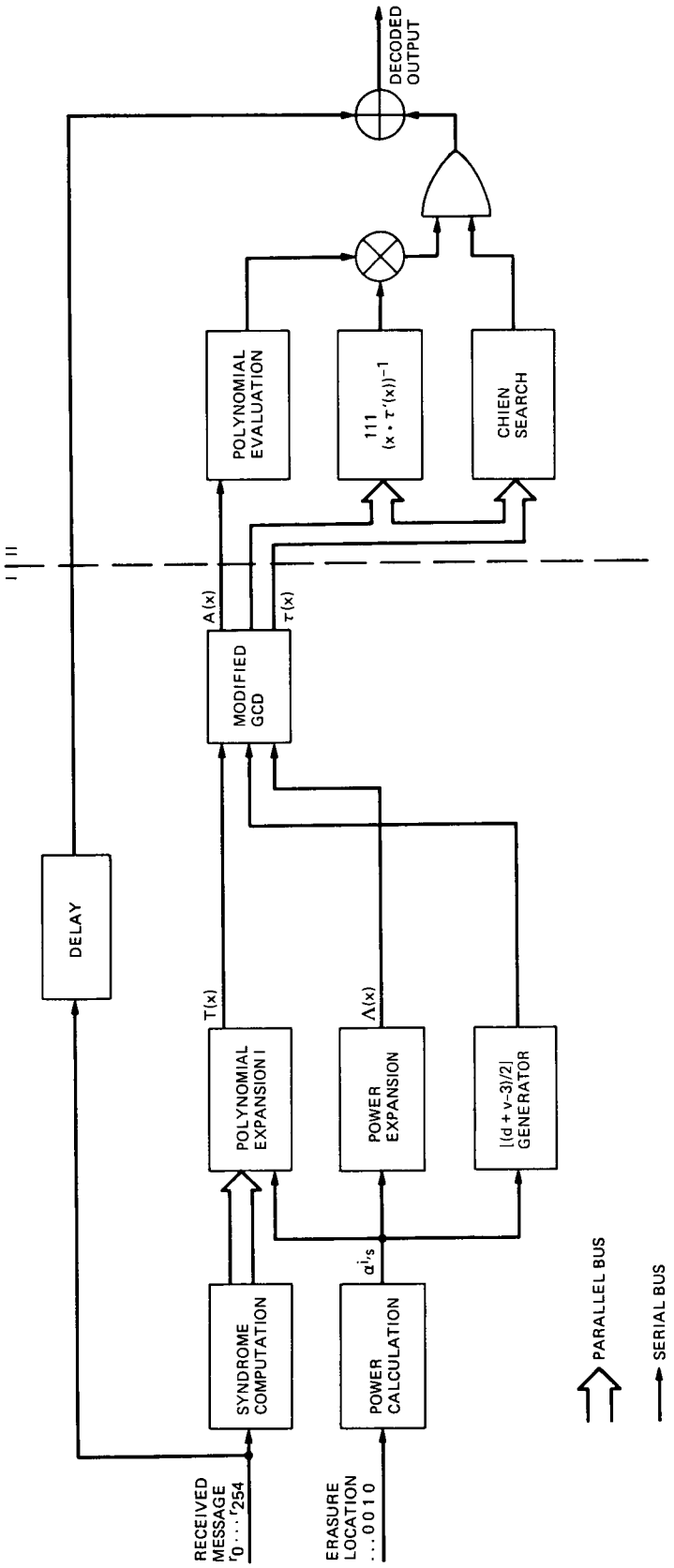


Fig. 2. A modified block diagram of a pipeline (255,223) RS time domain decoder

Appendix

Table A-1. Representations of the elements of $GF(2^4)$ generated by $\alpha^4 + \alpha + 1 = 0$

	α^3	α^2	α	α^0
α^0	0	0	0	1
α^1	0	0	1	0
α^2	0	1	0	0
α^3	1	0	0	0
α^4	0	0	1	1
α^5	0	1	1	0
α^6	1	1	0	0
α^7	1	0	1	1
α^8	0	1	0	1
α^9	1	0	1	0
α^{10}	0	1	1	1
α^{11}	1	1	1	0
α^{12}	1	1	1	1
α^{13}	1	1	0	1
α^{14}	1	0	0	1

More on the Decoder Error Probability for Reed-Solomon Codes

K.-M. Cheung
Communications Systems Research Section

This article is an extension of a recent paper by McEliece and Swanson dealing with the decoder error probability for Reed-Solomon codes (more generally, linear MDS codes). McEliece and Swanson offered an upper bound on $P_E(u)$*, the decoder error probability given that* u *symbol errors occur. This upper bound is slightly greater than* Q*, the probability that a completely random error pattern will cause decoder error. In this article, by using a combinatoric technique–the principle of inclusion and exclusion–an exact formula for* $P_E(u)$ *is derived.*

The $P_E(u)$*'s for the (255, 223) Reed-Solomon Code used by NASA, and for the (31,15) Reed-Solomon code (JTIDS code), are calculated using the exact formula, and the* $P_E(u)$*'s are observed to approach the* Q*'s of the codes rapidly as* u *gets large. An upper bound for the expression* $|[P_E(u)/Q] - 1|$ *is derived, and is shown to decrease nearly exponentially as* u *increases. This proves analytically that* $P_E(u)$ *indeed approaches* Q *as* u *becomes large, and some laws of large numbers come into play.*

I. Weight Distribution Formula for Decodable Words in a Linear MDS Code

A. Introduction

We begin with the following definitions. Let C be a linear code of length n, dimension k, and minimum distance d. Let q be a positive power of a prime. An (n,k,d) linear code C over $GF(q)$ is *maximum distance separable* (MDS) if the Singleton bound is achieved; that is, $d = n - k + 1$. A code is t-error correcting if for some integer t, $2t \leqslant d - 1$.

The class of Reed-Solomon (RS) codes is a subclass of MDS codes. Reed-Solomon codes are used in many sectors of today's industry. Some examples are the (255, 223) 16-error correcting RS code (the NASA code) in deep space communications, the (31,15) 8-error correcting RS code (the JTIDS code) in military communications, and the Cyclic Interleaving RS Code (CIRC) in the compact disc industry. A detailed treatment of MDS codes, their properties and open questions about them is given in [1]. The weight distribution of a linear MDS code with the parameters n, k, d, t, and q was independently found by three groups of researchers: Assmus, Mattson and Turyn [2], Forney [3], and Kasami, Lin and Peterson [4].

In Section I, we rederive the weight distribution formula for a linear MDS code by using the principle of inclusion and

exclusion, and then extend this method to obtain the exact weight distribution formula for "decodable words" in any linear MDS code. By decodable words, we mean all the words that lie within distance t from a codeword. If we assume the decoder to be a bounded distance decoder, then the weight distribution formula for the decodable words can be used to find the undetected error probability for linear MDS codes. This will be discussed in detail in Section II.

Section I is divided into 5 parts. Part I.A is a brief introduction. In I.B, we review some basic mathematical tools that are needed to derive the formulae. In I.C, we first derive the weight distribution formula for the number of codewords in a linear MDS code, and then we derive the weight distribution formula for the number of decodable words in a linear MDS code. In I.D, we give some numerical examples, and finally, in I.E, we end Section I of this article with some concluding remarks.

B. Some Basic Tools

In this part, we review the basic tools that are required to derive the weight distribution formulae for the number of codewords in a linear MDS code and for the number of decodable words in a linear MDS code.

Let C be an (n, k) code over $GF(q)$, not necessarily linear. If we examine any set of $k-1$ components of the codewords, we find that there are only q^{k-1} possibilities for the q^k codewords. Thus, there must be a pair of codewords that agree on these $k-1$ components, and so the minimum distance d of the code must satisfy $d \leqslant n-k+1$. This upper bound on d is known as the Singleton bound, and a code for which $d = n-k+1$ is called an MDS code. RS codes and cosets of RS codes are examples of MDS codes.

One important tool that we need is the basic combinatoric property of the MDS code. Let K be a subset of k coordinate positions of an MDS code. If two codewords were equal on K, the distance between them would be at most $n-k$. This contradicts the fact that $d = n-k+1$. Thus, all q^k codewords are different in K. Let $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$ be a k-tuple of elements from $GF(q)$. From the above argument, there exists a unique codeword whose k coordinates in K equal the k components of α. We call this important fact the basic combinatorial property of MDS codes.

Another important tool that we need is the principle of inclusion and exclusion [5]. Suppose we have N objects and a number of properties $P(1), \cdots, P(n)$. Let N_i be the number of objects with property $P(i)$, and $N_{i_1,i_2,\cdots,i_r}$ be the number of objects with properties $P(i_1), P(i_2), \cdots, P(i_r)$. The number of objects $N(0)$ with none of the properties is given by the following formula:

$$N(0) = N - \sum_{i} N_i + \sum_{i_1 < i_2} N_{i_1 i_2} + \cdots + (-1)^r \times \sum_{i_1 < i_2 \cdots < i_r} N_{i_1 i_2 \cdots i_r} + \cdots + (-1)^n N_{1\,2\,3 \cdots n} \tag{1}$$

The proof can be found in [5].

The basic combinatorial property of MDS codes and the principle of inclusion and exclusion will be referred to in the proofs in later sections.

C. Derivation of Formulae

This part is divided into three subparts. In the first, we derive the formula for the number of codewords of weight u in a linear MDS code, using the principle of inclusion and exclusion. In the second, we extend this idea by deriving a general formula for the number of decodable words of weight u. Last of all, in the third, we simplify the key formula by using some combinatoric identities.

1. Formula for the number of codewords of weight u. Let $\bar{c}$ be some codeword of C. Let $\bar{c}$ have a Hamming weight u, $u \geqslant d$. Let the coordinates of codeword $\bar{c}$ be indexed by $\{0, 1, 2, \cdots, n-1\}$. Define $v = n-u$. Then $\bar{c}$ has v zeros. We now want to find the number of codewords of weight u in C having exactly v zeros at some particular v coordinates where $v = n(u=0)$ or $v \leqslant n-d = k-1 (u \geqslant d)$. Since the code is linear, the number of codewords of weight zero ($u = 0$) is one—the all zero codeword. The following discussion applies only to codewords with weight $u \geqslant d$.

Let V be a set of v coordinates, $|V| = v$. Let $\{i_1, i_2, \cdots, i_j\} \subset \{1, 2, \cdots, n\} - V$ be a set of j coordinates. Define $S(i_1, i_2, \cdots, i_j) = \{\bar{c} : \bar{c} \in C$ and $\bar{c}$ has zeros in V and $\{i_1, i_2, \cdots, i_j\}\}$. For $j \leqslant k-v$, the number of zeros in a codeword in $S(i_1, i_2, \cdots, i_j)$ is at least $j+v \leqslant k (j+v \leqslant k)$. By using the basic combinatorial property of MDS code, for each particular choice of $\{i_1, i_2, \cdots, i_j\}$ we can specify q^{k-v-j} codewords having zeros at V and $\{i_1, i_2, \cdots, i_j\}$. So

$$|S(i_1, i_2, \cdots, i_j)| = q^{k-v-j} \qquad 0 \leqslant j \leqslant k-v \tag{2}$$

For $j \geqslant k-v+1$, the number of zeros in a codeword is $j+v \geqslant k+1$. This implies that the weight of the codeword is less than d, so $S(i_1, i_2, \cdots, i_j) = \{\bar{0}\}$. That is,

$$|S(i_1, i_2, \cdots, i_j)| = 1 \qquad k-v+1 \leqslant j \leqslant u \tag{3}$$

Note that we choose $i_1, i_2, \cdots, i_j$ from a set of $u = n - v$ coordinates so that for every choice of j, we have $\binom{u}{j} S(i_1, i_2, \cdots, i_j)$'s.

By the principle of inclusion and exclusion, the number of codewords with exactly v zeros at V equals

$$|S(0)| - \sum_{i_1} |S(i_1)| + \cdots + (-1)^u |S(i_1, i_2, \cdots, i_u)|$$

$$= \sum_{j=0}^{k-v-1} (-1)^j \binom{u}{j} q^{k-v-j} + \sum_{j=k-v}^{u} (-1)^j \binom{u}{j}$$

$$= \sum_{j=0}^{u-d} (-1)^j \binom{u}{j} q^{u-d-j+1} - \sum_{j=0}^{u-d} (-1)^j \binom{u}{j}$$

$$= \sum_{j=0}^{u-d} (-1)^j \binom{u}{j} (q^{u-d-j+1} - 1)$$

We have $\binom{n}{v} = \binom{n}{u}$ ways to choose v zeros from $\{0, 1, 2, \cdots, n-1\}$. Thus, the number of codewords of weight u, which is denoted by A_u, is given by the following expression:

$$A_u = \binom{n}{u} \sum_{j=0}^{u-d} (-1)^j \binom{u}{j} (q^{u-d-j+1} - 1) \qquad d \leqslant u \leqslant n \tag{4}$$

After deriving this relatively simple formula for the number of codewords of weight u in a linear MDS code, we proceed to derive the more complicated formula for the number of decodable words of weight u in a linear MDS code.

2. General formula for the number of decodable words of weight u. Let D be the set of decodable words in an MDS code. Let V be a set of v coordinates, $|V| = v$. Let $\{i_1, i_2, \cdots, i_j\}$ be a set of j coordinates, where $\{i_1, i_2, \cdots, i_j\} \subset \{0, 1, 2, \cdots, n-1\} - V$. Define $S(i_1, i_2, \cdots, i_j) = \{\bar{d} : \bar{d} \in D$ and $\bar{d}$ has zeros in V and $\{i_1, i_2, \cdots, i_j\}\}$. We proceed to derive the weight distribution formula for the number of decodable words of weight u in a linear MDS code by using the principle of inclusion and exclusion. Our problem is now reduced to finding the cardinality of $S(i_1, i_2, \cdots, i_j)$ for all j subjected to a given V. This problem is solved with the help of the following theorems.

Theorem 1:

$$|S(i_1, i_2, \cdots, i_j)| = q^{u-d+1-j} V_n(t) \qquad 0 \leqslant j \leqslant u - d \tag{5}$$

where

$$V_n(t) = \sum_{i=0}^{t} \binom{n}{i} (q-1)^i$$

Proof: The argument here is similar to the derivation given in I.C.1, above. We note that each coset of a linear MDS code is also an MDS code. Also, since all words lying within the Hamming spheres (with volume $V_n(t)$) that surround codewords are decodable words, we have $V_n(t)$ disjoint cosets that contain decodable words. From the basic combinatorial property of the MDS code we can, for each particular choice of $\{i_1, i_2, \cdots, i_j\}$, specify $q^{k-v-j} = q^{u-d+1-j}$ decodable words to each of these cosets. Thus, we have altogether $q^{u-d+1-j} V_n(t)$ decodable words having zeros at V and $\{i_1, i_2, \cdots, i_j\}$. This completes the proof. ∎

Theorem 2:

$$|S(i_1, i_2, \cdots, i_j)| = \sum_{w=d-u+j}^{t} \binom{n-u+j}{w}$$

$$\times \sum_{i=0}^{w-d+u-j} (-1)^i \binom{w}{i} (q^{w-d+u-j-i+1} - 1)$$

$$\times \sum_{s=w}^{t} \binom{u-j}{s-w} (q-1)^{s-w}$$

$$+ \sum_{i=0}^{t} \binom{u-j}{i} (q-1)^i \tag{6}$$

for $u - d + 1 \leqslant j \leqslant u - d + t$.

Proof: For $u - d + 1 = k - v \leqslant j$, the number of zeros in a decodable word is equal to $v + j \geqslant k$. Since $\bar{d}$ is a decodable word, $\bar{d}$ can be uniquely decomposed into a codeword $\bar{c}$ and an error pattern $\bar{e}$ with weight that is less than or equal to t. If we "project" $\bar{c}$ onto $V \cup \{i_1, i_2, \cdots, i_j\}$, then the result will be a certain $(v + j, k)$ code. Since the parent code has a minimum distance $d = n - k + 1$, the new code must have a minimum distance $d' \geqslant d - (n - v - j) = (v + j) - k + 1$. Since it is impossible for d' of the $(v + j, k)$ code to be greater than

$(v + j) - k + 1$ (because of the Singleton bound), d' must be equal to $d - (n - v - j) = (v + j) - k + 1$.

If $\bar{c} + \bar{e}$ vanishes on $V \cup \{i_1, i_2, \cdots, i_j\}$, then $\bar{c}$ must have weight that is less than or equal to t on $V \cup \{i_1, i_2, \cdots, i_j\}$. Let w be the weight of $\bar{c}$ on $V \cup \{i_1, i_2, \cdots, i_j\}$. From the above argument we also know that C, when restricted to $V \cup \{i_1, i_2, \cdots, i_j\}$, is a linear $(v + j, k)$ MDS code with a minimum distance $d - (n - v - j) = (v + j) - k + 1$. Thus, w is either 0 (in the case of the all-zero codeword) or between $d - u + j$ and t. So the number of codewords of weight w in the $(v + j, k)$ MDS code is (by using Eq. (4))

$$\binom{n-u+j}{w} \sum_{i=0}^{w-(d-(u-j))} (-1)^i \binom{w}{i} (q^{w-(d-(u-j))-i+1} - 1)$$

for $d - u + j \leqslant w \leqslant t$ and 1 for $w = 0$. For each codeword $\bar{c}$ with weight w in $V \cup \{i_1, i_2, \cdots, i_j\}$, where $d' \leqslant w \leqslant t$ $(d' = v + j - k + 1)$, we must count the number of $\bar{e}$'s such that $\bar{c} + \bar{e}$ vanishes on $V \cup \{i_1, i_2, \cdots, i_j\}$. Suppose that $\bar{e}$ has weight $s \geqslant w$. $\bar{e}$ must match $\bar{c}$ exactly on $V \cup \{i_1, i_2, \cdots, i_j\}$, but the $s - w$ other nonzero components can be arbitrarily placed outside $V \cup \{i_1, i_2, \cdots, i_j\}$. Then the total number of $\bar{e}$'s for a given $\bar{c}$ of weight w on $V \cup \{i_1, i_2, \cdots, i_j\}$ is

$$\sum_{s=w}^{t} \binom{u-j}{s-w} (q-1)^{s-w}$$

When $w = 0$, all components of $\bar{e}$ must lie outside the set $V \cup \{i_1, i_2, \cdots, i_j\}$. So there are

$$\sum_{i=0}^{t} \binom{u-j}{i} (q-1)^i$$

$\bar{e}$'s for the case $w = 0$. Combining the above results, we obtain the theorem. ∎

Theorem 3:

$$|S(i_1, i_2, \cdots, i_j)| = \sum_{i=0}^{t} \binom{u-j}{i} (q-1)^i \qquad (7)$$

for $u - d + t + 1 \leqslant j \leqslant u - t - 1$.

Proof: For $k - v + t \leqslant j \leqslant u - t - 1$, the number of zeros in a decodable word is greater than or equal to $k + t$ but less than or equal to $n - t - 1$. Thus any decodable words in $S(i_1, i_2, \cdots, i_j)$ have weight that is less than or equal to $d - t - 1$. It is not hard to see that the element of $S(i_1, i_2, \cdots, i_j)$ cannot be decoded into a codeword of weight other than $\bar{0}$. Therefore, $S(i_1, i_2, \cdots, i_j)$ contains all words having weight that is less than or equal to t in the coordinates $\{0, 1, \cdots, n-1\} - (V \cup \{i_1, i_2, \cdots, i_j\})$. This completes the proof. ∎

Theorem 4:

$$|S(i_1, i_2, \cdots, i_j)| = q^{u-j} \qquad \text{for } u - t \leqslant j \leqslant u \qquad (8)$$

Proof: Since j is greater than or equal to $u - t$, the number of zeros is equal to $v + j$ and is greater than or equal to $n - t$. Therefore, the number of nonzero components is less than or equal to t. Thus, all words with zeros on $V \cup \{i_1, i_2, \cdots, i_j\}$ are decodable and this completes the proof. ∎

As in I.C.1, we choose $i_1, i_2, \cdots, i_j$ from $v = n - u$ coordinates. Thus, for every choice of j, we have $\binom{u}{j} S(i_1, i_2, \cdots, i_j)$'s. Denote $N_j = \binom{u}{j} \mid S(i_1, i_2, \cdots, i_j) \mid$. Again, by the principle of inclusion and exclusion, we see that the number of decodable words which have exactly $v = n - u$ zeros at V equals

$$\sum_{j=0}^{u} (-1)^j N_j$$

However, we have $\binom{n}{u} = \binom{n}{v}$ ways to choose v zeros from 0, $1, \cdots, n - 1$. Thus, the number of decodable words of weight u is given by

$$D_u = \binom{n}{u} \sum_{j=0}^{u} (-1)^j N_j \qquad \text{for } d - t \leqslant u \leqslant n \qquad (9)$$

3. Simplification of the key formula. The weight enumerator formula that we have just derived is complicated and clumsy. There are four different expressions for N_j's, and these expressions are combined together by the inclusion and exclusion formula. The following theorem will show that the weight distribution formula for the number of decodable words in a linear MDS code can be simplified, and that there are only two expressions for the N_j's.

Theorem 5:

$$A = \sum_{j=0}^{u-t-1} (-1)^j \binom{u}{j} \sum_{i=0}^{t} \binom{u-j}{i} (q-1)^i + \sum_{j=u-t}^{u} (-1)^j \binom{u}{j} q^{u-j} = 0 \qquad (10)$$

Proof:

$$A = \sum_{j=0}^{u-t-1} (-1)^j \binom{u}{j} \left[q^{u-j} - \sum_{i=t+1}^{u-j} \binom{u-j}{i} (q-1)^i \right]$$

$$+ \sum_{j=u-t}^{u} (-1)^j \binom{u}{j} q^{u-j}$$

$$= \sum_{j=0}^{u} (-1)^j \binom{u}{j} q^{u-j}$$

$$- \sum_{j=0}^{u-t-1} (-1)^j \binom{u}{j} \sum_{i=t+1}^{u-j} \binom{u-j}{i} (q-1)^i$$

$$= (q-1)^u - \sum_{i=t+1}^{u} (q-1)^i \sum_{j=0}^{u-i} \binom{u-j}{i} \binom{u}{j} (-1)^j$$

$$= (q-1)^u - (q-1)^u$$

$$- \sum_{i=t+1}^{u-1} (q-1)^i \sum_{j=0}^{u-i} \binom{u-j}{i} \binom{u}{j} (-1)^j$$

Notice that

$$\binom{u}{j}\binom{u-j}{i} = \binom{u}{i}\binom{u-i}{j}$$

and

$$\sum_{j=0}^{u-i} \binom{u-i}{j} (-1)^j = 0$$

then

$$\sum_{j=0}^{u-i} \binom{u-j}{i}\binom{u}{j} (-1)^j = \binom{u}{i} \sum_{j=0}^{u-i} \binom{u-i}{j} (-1)^j = 0$$

for $t+1 \leqslant i \leqslant u-1$.

Thus, $A = 0$ and the theorem is proved. ∎

With Theorem 5 and Eqs. (5), (6), (7), (8), and (9), the weight enumerator formula can be simplified as follows:

$$D_u = \binom{n}{u} \sum_{j=0}^{u-d+t} (-1)^j N_j \tag{11}$$

for $d - t \leqslant u \leqslant n$

$$N_j = \binom{u}{j} \left[q^{u-d+1-j} V_n(t) - \sum_{i=0}^{t} \binom{u-j}{i} (q-1)^i \right] \tag{12}$$

for $0 \leqslant j \leqslant u - d$

$$N_j = \binom{u}{j} \left[\sum_{w=d-u+j}^{t} \binom{n-u+j}{w} \right.$$

$$\times \sum_{i=0}^{w-d+u-j} (-1)^i \binom{w}{i} (q^{w-d+u-j-i+1} - 1)$$

$$\left. \times \sum_{s=w}^{t} \binom{u-j}{s-w} (q-1)^{s-w} \right] \tag{13}$$

for $u - d + 1 \leqslant j \leqslant u - d + t$.

Examples will be found in Tables 1 and 2.

D. Remarks

The formula for the number of decodable words of weight u, where $d - t \leqslant u \leqslant n$, has been derived in the previous parts of this section. If we set $t = 0$, then we get back the weight enumerator for linear MDS code–Eq. (4). In the case of $u = d - t$, for example, we have

$$D_{d-t} = \binom{n}{d}\binom{d}{t}(q-1)$$

and the answer is consistent with the result derived in [6].

The formula is a bit clumsy, but can be easily implemented by computer program.

II. Decoder Error Probability of a Linear MDS Code

A. Number of Decodable Words vs. Decoder Error Probability

Let C be an (n, k, d) linear code capable of correcting t errors. When a codeword $\bar{c} \in C$ is transmitted over a com-

munication channel, channel noise may corrupt the transmitted signals. As a result, the receiver receives the corrupted version of the transmitted codeword $\bar{c} + \bar{e}$, where $\bar{e}$ is an error pattern of some weight u. If $u \leqslant t$, then a bounded distance decoder on the receiver's end detects and corrects the error $\bar{e}$ and recovers $\bar{c}$. If $u > t$, then the decoder fails and does one of two things:

(1) It detects the presence of the error pattern but is unable to correct it.

(2) It misinterprets (miscorrects) the received pattern $\bar{c} + \bar{e}$ for some other codeword $\bar{c}'$ if the received pattern falls into the radius t Hamming sphere of $\bar{c}'$.

Case (2) is, in most cases, more serious than case (1). This can occur (with a nonzero probability) when an error pattern $\bar{e}$ is of weight $u \geqslant d - t$. Let us further assume that all error patterns of weight u are equally probable, and let us use $P_E(u)$ [7] to denote the decoder error probability given that an error pattern of weight u occurs. It is not hard to see that $P_E(u)$ is given by the following expression:

$$P_E(u) = \frac{D_u}{\binom{n}{u}(q-1)^u} \qquad \text{for } d - t \leqslant u \leqslant n \quad (14)$$

That is, $P_E(u)$ is the ratio of the number of decodable words of weight u to the number of words of weight u in the whole vector space. Thus, the problem of finding the $P_E(u)$'s is essentially the same as the problem of finding the weight distribution of the set of decodable words. Equations (11), (12) and (13) of Section I and Eq. (14) of Section II together enable us to find the exact decoder error probability of a linear MDS code.

Let the probability that a completely random error pattern will cause decoder error be denoted by Q. It is the ratio of the number of decodable words to the cardinality of the whole vector space. That is,

$$Q = \frac{(q^k - 1)\, V_n(t)}{q^n} \cong q^{-r}\, V_n(t) \qquad (15)$$

where $r = n - k$ is the code's redundancy and

$$V_n(t) = \sum_{i=0}^{t} \binom{n}{i} (q-1)^i$$

is the volume of a Hamming sphere of radius t. It is shown in the next part of this section that if $q \geqslant n$, which is generally true, then $P_E(u)$ approaches Q very rapidly as u increases.

B. Examples and Observations

Two well-known examples of linear MDS codes—the NASA code and the JTIDS code—are tabulated in Table 3 and Table 4, respectively. In these two examples, we observe that $P_E(u)$ approaches the constant Q as u increases. In fact, $P_E(u)$ approaches Q rapidly for $u \ll n$. In the case of large q and $q \geqslant n$, $P_E(u)$ approaches Q even for $u < d$. The $P_E(u)$ and Q of the NASA code agree to eight significant digits for $u \geqslant 26$ ($d = 33$). If $P_E(u)$ and Q are interpreted combinatorically as ratios, then we have the following relationship:

$$\frac{\#\text{ of decodable words of weight } u}{\#\text{ of vectors of weight } u} \rightarrow \frac{\#\text{ of decodable words}}{\#\text{ of words in vector space}}$$

This astonishing relationship cited above implies that a linear MDS code, which possesses rigid algebraic and combinatoric structures, behaves (in some sense) like a random code with no structure at all. Some laws of large number come into play somehow.

In order to describe analytically how fast $P_E(u)$ approaches Q when u is large, an upper bound on the expression $|[P_E(u)/Q] - 1|$ is derived in the following paragraphs. This upper bound is denoted by $\mathrm{U}(u)$, where $u \geqslant d$. It will be shown that $\mathrm{U}(u)$ approaches a very small number ϵ as u increases.

As in Section I, let D_u denote the exact number of decodable words of weight u. Let N_j's be the corresponding terms in the inclusion and exclusion formula of D_u as expressed in Eqs. (11), (12), and (13) of Section I. Let $\hat{D}_u$ denote the estimated number of decodable words of weight u. Let $\hat{N}_j$'s be the corresponding terms in the inclusion and exclusion formula of $\hat{D}_u$. The expression of $\hat{N}_j$, $0 \leqslant j \leqslant u$ is constructed by extrapolating the first term on the right-hand side of Eq. (12) of Section I from $0 \leqslant j \leqslant u - d$ to $0 \leqslant j \leqslant u$. We now have the following equations for $\hat{D}_u$ and $\hat{N}_j$:

$$\hat{D}_u = \binom{n}{u} \sum_{j=0}^{u} (-1)^j \hat{N}_j \qquad d - t \leqslant u \leqslant n \quad (16)$$

$$\hat{N}_j = \binom{u}{j} q^{u-d+1-j}\, V_n(t) \qquad 0 \leqslant j \leqslant u \quad (17)$$

Now we want to find an upper bound, denoted by U_j, for N_j in Eq. (13) of Section I for $u - d + 1 \leqslant j \leqslant u - d + t$.

$$N_j = \binom{u}{j} \sum_{w=d-u+j}^{t} \binom{n-u+j}{w}$$

$$\times \sum_{i=0}^{w-d+u-j} (-1)^i \binom{w}{i} (q^{w-d+u-j-i+1} - 1)$$

$$\times \sum_{s=w}^{t} \binom{u-j}{s-w} (q-1)^{s-w}$$

$$= \binom{u}{j} \sum_{w=d-u+j}^{t} \binom{n-u+j}{w} (q-1)$$

$$\times \left[\sum_{i=0}^{w-d+u-j} (-1)^i \binom{w-1}{i} q^{w-d+u-j-i} \right]$$

$$\times \sum_{s=w}^{t} \binom{u-j}{s-w} (q-1)^{s-w}$$

$$\leqslant \binom{u}{j} \sum_{w=d-u+j}^{t} \binom{n-u+j}{w} (q-1)\, q^{-d+u-j}\, q^w$$

$$\times \sum_{s=w}^{t} \binom{u-j}{s-w} (q-1)^{s-w}$$

$$= \binom{u}{j} \sum_{w=d-u+j}^{t} \binom{n-u+j}{w} q^{u-j-d+1} \left(\frac{q}{q-1}\right)^{w-1} (q-1)^w$$

$$\times \sum_{s=w}^{t} \binom{u-j}{s-w} (q-1)^{s-w}$$

$$\leqslant q^{u-j-d+1} \left(\frac{q}{q-1}\right)^{t-1} \binom{u}{j} \sum_{w=d-u+j}^{t} \binom{n-u+j}{w} (q-1)^w$$

$$\times \sum_{s=w}^{t} \binom{u-j}{s-w} (q-1)^{s-w}$$

$$= q^{u-j-d+1} \left(\frac{q}{q-1}\right)^{t-1} \binom{u}{j}$$

$$\times \sum_{s=d-u+j}^{t} (q-1)^s \sum_{w=d-u+j}^{s} \binom{n-u+j}{w} \binom{u-j}{s-w}$$

$$< q^{u-j-d+1} \left(\frac{q}{q-1}\right)^{t-1} \binom{u}{j} V_n(t)$$

$$= \left(\frac{q}{q-1}\right)^{t-1} \hat{N}_j \stackrel{\text{def}}{=} U_j$$

Note that $\binom{u}{j} q^{u-d+1-j} V_n(t) = \hat{N}_j < U_j$, and so $U_j \geqslant \max \{N_j, \hat{N}_j\}$. Also, with the additional assumption that $q \geqslant n$, which is generally true, U_j is a descending function of j.

Now let us consider the second term on the right-hand side of Eq. (12) of Section I, and denote it by $\Theta(u)$. We want to find an upper bound for $\Theta(u)$.

$$\Theta(u) = \sum_{j=0}^{u-d} (-1)^j \binom{u}{j} \sum_{i=0}^{t} \binom{u-j}{i} (q-1)^i$$

$$= \sum_{j=0}^{u-d} (-1)^j \sum_{i=0}^{t} \binom{u}{i} \binom{u-i}{j} (q-1)^i$$

$$= \sum_{i=0}^{t} \binom{u}{i} (q-1)^i \sum_{j=0}^{u-d} (-1)^j \binom{u-i}{j}$$

$$\leqslant \sum_{i=0}^{t} \binom{u}{i} (q-1)^i \sum_{j=0}^{u-d} \binom{u-i}{j}$$

$$\leqslant \sum_{i=0}^{t} \binom{u}{i} (q-1)^i \sum_{j=0}^{u-i} \binom{u-i}{j}$$

$$= \sum_{i=0}^{t} \binom{u}{i} (q-1)^i \, 2^{u-i}$$

$$= 2^u \sum_{i=0}^{t} \binom{u}{i} \left(\frac{q-1}{2}\right)^i$$

$$= 2^u V_u^*(t)$$

where

$$V_u^*(t) = \sum_{i=0}^{t} \binom{u}{i} \left(\frac{q-1}{2}\right)^i$$

We then want to find an upper bound of $|D_u - \hat{D}_u|$, where $d \leqslant u \leqslant n$. We have

$$|D_u - \hat{D}_u| = \left| \binom{n}{u} \left[\sum_{j=0}^{u-d+t} (-1)^j N_j - \sum_{j=0}^{u} (-1)^j \hat{N}_j \right] \right|$$

$$\leqslant \binom{n}{u} \left[\left| \sum_{j=u-d+1}^{u-d+t} (-1)^j N_j - \sum_{j=u-d+1}^{u} (-1)^j \hat{N}_j \right| + \Theta(t) \right]$$

$$= \binom{n}{u} \left[\left| \sum_{j=u-d+1}^{u} (-1)^j (N_j - \hat{N}_j) \right| + \Theta(u) \right]$$

(set $N_j = 0$ for $u - d + t + 1 \leqslant j \leqslant u$)

$$\leqslant \binom{n}{u} \left[\sum_{j=u-d+1}^{u} \left| N_j - \hat{N}_j \right| + \Theta(u) \right]$$

$$\leqslant \binom{n}{u} \left[\sum_{j=u-d+1}^{u} \max \{N_j, \hat{N}_j\} + \Theta(u) \right]$$

$$\leqslant \binom{n}{u} \left[\sum_{j=u-d+1}^{u} U_j + \Theta(u) \right]$$

$$\leqslant \binom{n}{u} [dU_{u-d+1} + \Theta(u)]$$

(U_j is a descending function)

$$= \binom{n}{u} \left[d \left(\frac{q}{q-1} \right)^{t-1} \binom{u}{d-1} V_n(t) + 2^u V_u^*(t) \right]$$

We are finally ready to derive an upper bound for $|[P_E(u)/Q] - 1|$. By the definition of $\hat{D}_u$ in Eqs. (16) and (17), it is not hard to see that

$$\hat{D}_u = \binom{n}{u} \sum_{j=0}^{u} (-1)^j \binom{u}{j} q^{u-d+1-j} V_n(t)$$

$$= \binom{n}{u} V_n(t) q^{-d+1} (q-1)^u$$

Now for $d \leqslant u \leqslant n$,

$$\left| \frac{P_E(u)}{Q} - 1 \right| = \left| \frac{q^n D_u}{\binom{n}{u} (q-1)^u q^k V_n(t)} - 1 \right|$$

$$= \left| \frac{q^n (D_u - \hat{D}_u)}{\binom{n}{u} (q-1)^u q^k V_n(t)} \right|$$

$$= \frac{q^n |D_u - \hat{D}_u|}{\binom{n}{u} (q-1)^u q^k V_n(t)}$$

$$\leqslant \left(\frac{q}{q-1} \right)^{t-1} \frac{q^{d-1} \binom{u}{d-1} d}{(q-1)^u} + \frac{q^{d-1} 2^u}{(q-1)^u} \frac{V_u^*(t)}{V_n(t)} \stackrel{\text{def}}{=} \mathrm{U}(u)$$

where

$$V_n(t) = \sum_{i=0}^{t} \binom{n}{i} (q-1)^i$$

and

$$V_u^*(t) = \sum_{i=0}^{t} \binom{u}{i} \left(\frac{q-1}{2}\right)^i$$

Thus, the upper bound $\mathbf{U}(u)$ of $|[P_E(u)/Q] - 1|$, which is a function of u for $d \leqslant u \leqslant n$, is given by the following equation:

$$\left|\frac{P_E(u)}{Q} - 1\right| \leqslant \left(\frac{q}{q-1}\right)^{t-1} \frac{q^{d-1}\binom{u}{d-1} d}{(q-1)^u} + \frac{q^{d-1}\, 2^u}{(q-1)^u} \frac{V_u^*(t)}{V_n(t)} = \mathbf{U}(u)$$

The upper bounds of $|[P_E(u)/Q] - 1|$ of the NASA code and the JTIDS code are tabulated in Table 5 and Table 6, respectively.

C. Remarks

With the assumptions that q is greater than or equal to n and that u is large compared to d, Eq. (18) shows that the upper bound of $|[P_E(u)/Q] - 1|$ is dominated by the denominator term $(q-1)^u$. Thus, the upper bound of $|[P_E(u)/Q] - 1|$ decays nearly exponentially as a function of u. This upper bound is not a very tight bound, but it is sufficient to illustrate the point that $P_E(u)$ approaches Q very rapidly as u increases.

References

[1] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, Amsterdam, The Netherlands: North-Holland, 1983.

[2] E. F. Assmus, H. F. Mattson, Jr., and R. J. Turyn, *Cyclic Codes*, AFCRL-65-332, Air Force Cambridge Research Labs, Bedford, Massachusetts, 1965.

[3] G. D. Forney, Jr., *Concatenated Codes*, Cambridge, Massachusetts: The MIT Press, 1966.

[4] T. Kasami, S. Lin, and W. W. Peterson, "Some Results on Weight Distributions of BCH codes," *IEEE Trans. Inform. Theory*, vol. IT-12, p. 274, April 1966.

[5] M. Hall, *Combinatorial Theory*, Waltham, Massachusetts: Blaisdell, 1967.

[6] E. R. Berlekamp and J. L. Ramsey, "Readable Erasures Improve the Performance of Reed-Solomon Codes," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 632–633, September 1978.

[7] R. J. McEliece and L. Swanson, "On the Decoder Error Probability for Reed-Solomon Codes," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 701–703, September 1986.

Table 1. (4,2) MDS code over *GF*(5) with $t = 1$

Weight	Number of decodable words	Upper bound [6]
0	1	–
1	16	–
2	48	48
3	192	272
4	168	272

Total number of decodable words = $q^k V_n(t) = 425$.

Table 2. (6,3) MDS code over *GF*(4) with $t = 1$

Weight	Number of decodable words	Upper bound [6]
0	1	–
1	18	–
2	0	–
3	180	180
4	405	855
5	378	1026
6	234	513

Total number of decodable words = $q^k V_n(t) = 1216$.

Table 3. NASA Code: (255,223); RS code: $q = 256$, $t = 16$

$$P_E(17) = 9.4641648 \times 10^{-15}$$
$$P_E(18) = 1.9130119 \times 10^{-14}$$
$$P_E(19) = 2.4010995 \times 10^{-14}$$
$$P_E(20) = 2.6598044 \times 10^{-14}$$
$$P_E(21) = 2.6017177 \times 10^{-14}$$
$$P_E(22) = 2.6076401 \times 10^{-14}$$
$$P_E(23) = 2.6087596 \times 10^{-14}$$
$$P_E(24) = 2.6088773 \times 10^{-14}$$
$$P_E(25) = 2.6088880 \times 10^{-14}$$
$$P_E(26) = 2.6088888 \times 10^{-14}$$
$$P_E(27) = 2.6088888 \times 10^{-14}$$
$$P_E(28) = 2.6088888 \times 10^{-14}$$
$$P_E(29) = 2.6088888 \times 10^{-14}$$
$$P_E(30) = 2.6088888 \times 10^{-14}$$
$$\vdots \qquad \vdots$$

Table 4. JTIDS Code: (31,15); RS code: $q = 32$, $t = 8$

$P_E(9) = 3.7493431 \times 10^{-7}$
$P_E(10) = 1.4392257 \times 10^{-6}$
$P_E(11) = 2.9507015 \times 10^{-6}$
$P_E(12) = 4.3287703 \times 10^{-6}$
$P_E(13) = 5.1888955 \times 10^{-6}$
$P_E(14) = 5.5466000 \times 10^{-6}$
$P_E(15) = 5.6291887 \times 10^{-6}$
$P_E(16) = 5.6296979 \times 10^{-6}$
$P_E(17) = 5.6255686 \times 10^{-6}$
$P_E(18) = 5.6256673 \times 10^{-6}$
$P_E(19) = 5.6259065 \times 10^{-6}$
$P_E(20) = 5.6258313 \times 10^{-6}$
$P_E(21) = 5.6258455 \times 10^{-6}$
$P_E(22) = 5.6258434 \times 10^{-6}$
$P_E(23) = 5.6258437 \times 10^{-6}$
$P_E(24) = 5.6258437 \times 10^{-6}$
.
.
.

Table 5. NASA Code: (255,223); RS code: $q = 256$, $t = 16$

u	U(u)
33	5.133
34	0.3422
35	0.0157
36	5.526×10^{-4}
37	1.512×10^{-5}
.	.
.	.
.	.

Table 6. JTIDS Code: (31,15); RS code: $q = 32$, $t = 8$

u	U(u)
17	19.35
18	5.618
19	1.148
20	0.1851
21	0.02508
.	.
.	.
.	.

On the VLSI Design of a Pipeline Reed–Solomon Decoder Using Systolic Arrays

H. M. Shao and L. J. Deutsch
Communications Systems Research Section
I. S. Reed
University of Southern California

A new VLSI design of a pipeline Reed-Solomon decoder is presented. The transform decoding technique used in a previous article is replaced by a time domain algorithm through a detailed comparison of their VLSI implementations. A new architecture that implements the time domain algorithm permits efficient pipeline processing with reduced circuitry. Erasure correction capability is also incorporated with little additional complexity. By using a multiplexing technique, a new implementation of Euclid's algorithm maintains the throughput rate with less circuitry. Such improvements result in both enhanced capability and significant reduction in silicon area.

I. Introduction

Recently a VLSI design of a pipeline Reed-Solomon decoder was presented [1]. A modified form of Euclid's algorithm was developed which avoided computations of inverse elements. A systolic array architecture was designed, from a suggestion by Brent and Kung [2], to implement the modified Euclid's algorithm. More recently, another VLSI design of an RS decoder was introduced [3]. It combined the algorithm in [4] and the modified Euclid's algorithm instead of the continued fraction technique. The decoder design in [3] used a time domain decoding algorithm to reduce the massive circuitry required by the inverse transform in [1]. The decoder design also included the erasure correction capability, and, during the design process, a recursive architecture was derived to implement the modified Euclid's algorithm by far fewer circuits than used in [1].

It has been pointed out [5] that the errata locator polynomial can be obtained directly from the Massey-Berlekemp algorithm if initialized properly. This suggestion led to improvements in the VLSI design in [3].

In this article, an efficient time domain RS decoding algorithm is described and verified. It is shown that the modified Euclid's algorithm can produce the errata locator polynomial and errata evaluator polynomial simultaneously, similar to the Massey-Berlekemp algorithm. The VLSI architectures for syndrome computations, polynomial expansions, modified Euclid's algorithm performance, and polynomial evaluations are also described.

This work was carried out during the architectural phase of the Advanced Reed-Solomon Decoder (ARSD) project and

should be viewed as a companion to the recent work of Truong, *et al.* [6]. In that article, a transform domain decoder architecture is developed which, due to its design simplicity, has been chosen for the prototype VLSI implementation of the ARSD. However, the work presented here and in [6] clearly shows that the time domain architecture has many desirable features which make it an attractive candidate for future VLSI implementation.

II. The Time Domain Reed–Solomon Decoding Algorithm

Let $N = 2^m - 1$ be the length of the (N, I) RS code with design distance d.

Let

$$R(X) = \sum_{i=0}^{N-1} r_i X^i = r_{N-1} X^{N-1} + \ldots + r_1 X + r_0$$

be the received message. Suppose e errors and E erasures occur, and $2e + E < d - 1$. Define $\Lambda = \{\alpha^{-i} | r_i$ declared as an erasure$\}$.

The decoding algorithm is as follows:

Step 1. Compute the syndromes

$$S_k = \sum_{i=0}^{N-1} r_i X^i \Big|_{X=\alpha^k}$$

$$= \sum_{i=0}^{N-1} r_i \alpha^{ki} \qquad \text{for } 1 \leqslant k \leqslant d-1 \tag{1}$$

Form a syndrome polynomial

$$S(X) = \sum_{k=1}^{d-1} S_k X^{k-1} \tag{2}$$

Step 2. Compute the erasure locator polynomial $\Lambda(X)$. Assume the erasure location information is received in the form of a binary sequence synchronous to the received message

$$R(X) = \sum_{i=0}^{N-1} r_i X^i$$

Then for each symbol r_i that is labeled as an erasure, α^{-i} should be the root of the erasure locator polynomial $\Lambda(X)$. That is,

$$\Lambda(X) = \prod_{\alpha^{-i} \in \Lambda} (X - \alpha^{-i}) \tag{3}$$

Step 3. Multiply the syndrome polynomial $S(X)$ by the erasure locator polynomial $\Lambda(X)$ to form the modified syndrome polynomial

$$T(X) = S(X)\,\Lambda(X) \bmod X^{d-1}$$

$$= \sum_{k=1}^{d-1} T_k X^{k-1} \tag{4}$$

Step 4. If $\deg(\Lambda(X)) > \deg(T(X))$, then no error has occurred, i.e., $e = 0$. Thus there is no need to perform the modified Euclid's algorithm. Let the errata locator polynomial $\sigma(X) = \Lambda(X)$ and the errata evaluator polynomial $\omega(X) = T(X)$. If $\deg(\Lambda(X)) \leqslant \deg(T(X))$, then perform a modified Euclid's algorithm on X^{d-1} and $T(X)$ with the following initializations:

$$\begin{aligned} \mu_0(X) &= \Lambda(X) & R_0(X) &= X^{d-1} \\ \lambda_0(X) &= 0 & Q_0(X) &= T(X) \end{aligned} \tag{5}$$

Compute the following iterations:

$$\begin{aligned} R_i(X) = {} & [\sigma_{i-1} b_{i-1} R_{i-1}(X) + \bar{\sigma}_{i-1} a_{i-1} Q_{i-1}(X)] \\ & - X^{|\varrho_{i-1}|} [\sigma_{i-1} a_{i-1} Q_{i-1}(X) \\ & + \bar{\sigma}_{i-1} b_{i-1} R_{i-1}(X)] \end{aligned} \tag{6}$$

$$\begin{aligned} \lambda_i(X) = {} & [\sigma_{i-1} b_{i-1} \lambda_{i-1}(X) + \bar{\sigma}_{i-1} a_{i-1} \mu_{i-1}(X)] \\ & - X^{|\varrho_{i-1}|} [\sigma_{i-1} a_{i-1} \mu_{i-1}(X) \\ & + \bar{\sigma}_{i-1} b_{i-1} \lambda_{i-1}(X)] \end{aligned} \tag{7}$$

$$Q_i(X) = \sigma_{i-1} Q_{i-1}(X) + \bar{\sigma}_{i-1} R_{i-1}(X) \tag{8}$$

$$\mu_i(X) = \sigma_{i-1} \mu_{i-1}(X) + \bar{\sigma}_{i-1} \lambda_{i-1}(X) \tag{9}$$

where a_{i-1} and b_{i-1} are the leading coefficients of $R_{i-1}(X)$ and $Q_{i-1}(X)$, respectively,

$$\ell_{i-1} = \deg(R_{i-1}(X)) - \deg(Q_{i-1}(X))$$

and

$$\begin{aligned} \sigma_{i-1} &= 1 \qquad \text{if } \ell_{i-1} \geqslant 0 \\ \sigma_{i-1} &= 0 \qquad \text{if } \ell_{i-1} < 0 \end{aligned} \tag{10}$$

Stop the iterations when $\deg(\lambda_i(X)) > \deg(R_i(X))$. Let the errata locator polynomial $\sigma(X) = \lambda_i(X)$ and the errata evaluator polynomial $\omega(X) = R_i(X)$. The $\sigma(X)$ and $\omega(X)$ polynomials, obtained by the modified Euclid's algorithm, both carry a common scale factor compared to those computed by the conventional Euclid's algorithm. But this scale factor does not affect the errata location computations or the errata magnitude computations.

Step 5. Evaluate the errata locator polynomial $\sigma(X)$ for α^{-i}, $i = 0, \ldots, N-1$ to find the roots of $\sigma(X)$. If $\sigma(\alpha^{-i}) = 0$, then r_i is a corrupted symbol.

Step 6. Compute the corresponding errata magnitudes by evaluating $\omega(X)$ and $\sigma'(X)$ for α^{-i}, $i = 0, \ldots, N-1$. That is, the errata magnitude

$$\hat{e}_i = -\frac{\omega(\alpha^{-i})}{\sigma'(\alpha^{-i})} \qquad 0 \leqslant i \leqslant N-1 \tag{11}$$

Note that the scale factor carried by $\omega(X)$ and $\sigma(X)$ is automatically cancelled by this division.

Step 7. Subtracting $\hat{e}_i$ from r_i yields the decoded codeword

$$\hat{C}_i = r_i - \hat{e}_i \qquad 0 \leqslant i \leqslant N-1 \tag{12}$$

Note that the modified Euclid's algorithm in Step 4 is a combination of three techniques. First, observe that the error locator polynomial $\lambda(X)$ and the errata evaluator polynomial $\omega(X)$ can be obtained from Euclid's algorithm by computing the GCD of the modified syndrome $T(X)$ and X^{d-1} with the following initializations:

$$\begin{aligned} \mu_0(X) &= 1 \qquad & R_0(X) &= X^{d-1} \\ \lambda_0(X) &= 0 \qquad & Q_0(X) &= T(X) \end{aligned} \tag{13}$$

Since e errors and E erasures occur and $2e + E \leqslant d-1$, as in Theorem 8.4 of [7], the following properties hold:

$$\deg(\lambda(X)) = e \qquad \deg(\omega(X)) < e + E \tag{14}$$

$$\text{GCD}(\lambda(X), \omega(X)) = 1 \tag{15}$$

$$e_i = -\frac{\omega(X)}{[\Lambda(X)\lambda(X)]'}\bigg|_{X=\alpha^{-i}} \tag{16}$$

$$\lambda(X)\,\Lambda(X)S(X) \equiv \omega(X) \bmod X^{d-1} \tag{17}$$

Applying properties (14) and (17) to Theorem 8.5 of [7] implies that there exist a unique j and a unique polynomial $\beta(X)$ such that

$$\lambda(X) = \beta(X)\,\lambda_j(X)$$

$$\omega(X) = \beta(X)\,R_j(X)$$

By properties (15) and (16), $\beta(X)$ is a constant, which can be taken to be unity without affecting the roots of $\lambda(X)$ or the magnitudes e_i. The second technique applied to the modified Euclid's algorithm is that the errata locator polynomial $\sigma(X) = \Lambda(X)\,\lambda(X)$ can be obtained directly from the Euclid's algorithm. To achieve this, $\mu_o(X)$ must be initialized to be the erasure locator polynomial $\Lambda(X)$ instead of 1, and the iteration stop criterion must be changed to $\deg(R_i(X)) < \deg(\lambda_i(X))$. Such a change simply results in all $\lambda_i(X)$ carrying the factor $\Lambda(X)$. The errata evaluator polynomial $\omega(X)$ is not affected by such initialization because $\lambda_i(X)$ does not involve the computation of $R_i(X)$. As will be shown later, using the modified Euclid's algorithm to compute the errata locator polynomial directly eliminates the need for polynomial multiplication circuits and delay lines in a VLSI pipeline implementation. Thirdly, the modified Euclid's algorithm uses cross multiplication and subtraction to replace polynomial division. Such operations eliminate the need to compute finite field inverse elements, which is performed by a table look-up, in this step. Since a look-up table involves the use of a large silicon area in VLSI, it is preferable to do this as infrequently as possible.

Example. Consider an RS (8, 4) code over GF (17) with generator polynomial $g(X) = (X-2)(X-2^2)(X-2^3)(X-2^4)$. Suppose two erasures and one error have occurred and the all zero codeword was sent. Let $R(X) = -2X^5 - 3X^2 + 2X$ be the received vector with locations X^5 and X^2 flagged as erasures. Thus the erasure locator polynomial

$$\Lambda(X) = (X - 2^{-5})(X - 2^{-2})$$

$$= X^2 + 13X + 2$$

(1) Compute the syndromes

$$S_k = \sum_{i=0}^{7} r_i 2^{ik} \qquad k = 1, 2, 3, 4$$

$$S_1 = R(2^1) = 13$$

$$S_2 = R(2^2) = 3$$

$$S_3 = R(2^3) = 10$$

$$S_4 = R(2^4) = 14$$

Form the syndrome polynomial

$$S(X) = \sum_{k=1}^{4} S_k X^{k-1} = S_4 X^3 + S_3 X^2 + S_2 X^1 + S_1$$

$$= 14X^3 + 10X^2 + 3X + 13$$

(2) Compute the modified syndromes

$$T(X) = S(X)\,\Lambda(X) \bmod X^4$$

$$= (14X^3 + 10X^2 + 3X + 13) \times (X^2 + 13X + 2) \bmod X^4$$

$$= 8X^3 + 4X^2 + 5X + 9$$

Thus

$$T_4 = 8, T_3 = 4, T_2 = 5, T_1 = 9$$

(3) Perform the modified Euclid's algorithm

$$\mu_0(X) = \Lambda(X) = X^2 + 13X + 2$$

$$\lambda_0(X) = 0$$

$$R_0(X) = X^4$$

$$Q_0(X) = T(X) = 8X^3 + 4X^2 + 5X + 9$$

$$R_1(X) = 8R_0(X) - XQ_0(X)$$

$$= 8X^4 - X(8X^3 + 4X^2 + 5X + 9)$$

$$= -4X^3 - 5X^2 - 9X$$

$$\lambda_1(X) = 8\lambda_0(X) - X\mu_0(X) = -X(X^2 + 13X + 2)$$

$$= -X^3 - 13X^2 - 2X$$

$$Q_1(X) = Q_0(X) = 8X^3 + 4X^2 + 5X + 9$$

$$\mu_1(X) = \mu(X) = X^2 + 13X + 2$$

$$R_2(X) = 8R_1(X) - (-4)Q_1(X)$$

$$= 8(-4X^3 - 5X^2 - 9X) + 4(8X^3 + 4X^2 + 5X + 9)$$

$$= 10X^2 - X + 2$$

$$\lambda_2(X) = 8\lambda_1(X) - (-4)\mu_1(X) = 8(X^3 - 13X^2 - 2X) + 4(X^2 + 13X + 2)$$

$$= 9X^3 + 2X^2 + 2X + 8$$

Since $\deg(\lambda_2(X)) - \deg(R_2(X)) = 1$, Stop.

Thus the errata evaluator is

$$\omega(X) = R_2(X) = 10X^2 - X + 2$$

and the errata locator is

$$\sigma(X) = \lambda_2(X) = 9X^3 + 2X^2 + 2X + 8$$

(4) Perform Chien search on $\sigma(X)$ and evaluate $-\omega(X)/\sigma'(X)$

$\sigma(2^{-7}) = 7;$ $\quad \hat{e}_7 = 0$

$\sigma(2^{-6}) = 12;$ $\quad \hat{e}_6 = 0$

$\sigma(2^{-5}) = 0;$ $\quad \hat{e}_5 = -\dfrac{\omega(2^{-5})}{\sigma'(2^{-5})} = -2$

$\sigma(2^{-4}) = 16;$ $\quad \hat{e}_4 = 0$

$\sigma(2^{-3}) = 8;$ $\quad \hat{e}_3 = 0$

$\sigma(2^{-2}) = 0;$ $\quad \hat{e}_2 = -\dfrac{\omega(2^{-2})}{\sigma'(2^{-2})} = -3$

$\sigma(2^{-1}) = 0;$ $\quad \hat{e}_1 = -\dfrac{\omega(2^{-1})}{\sigma'(2^{-1})} = 2$

$\sigma(2^{-0}) = 4;$ $\quad \hat{e}_0 = 0$

(5) $C_i = r_i - \hat{e}_i \qquad i = 7, 6, 5, 4, 3, 2, 1, 0$

$= (0, 0, -2, 0, 0, -3, 2, 0) - (0, 0, -2, 0, 0, -3, 2, 0)$

$= (0, 0, 0, 0, 0, 0, 0, 0)$

The VLSI architecture of the pipeline RS decoder is shown in Fig. 1. The syndromes $S(X)$ are computed by a form of polynomial evaluation. The α^k generation block converts binary erasure location information to powers of α which are the roots of the erasure locator polynomial. The modified syndromes $T(X)$ and the erasure locator polynomial $\Lambda(X)$ can be computed by two polynomial multiplication circuits. By the use of a multiplexing and recursive technique, the modified Euclid's algorithm is implemented with a significant reduction of cells over a previous design [1]. The errata evaluator polynomial $\omega(X)$ and the errata locator polynomial $\sigma(X)$ are then evaluated using two polynomial evaluation circuits different from the one used for syndrome computation. The errata locations thus obtained direct the subtractions of the errata from the received messages to produce the decoded messages. In the following, the VLSI design of each functional block is described.

III. VLSI Implementation of the Syndrome Computation

The syndrome computation

$$S_k = \sum_{i=0}^{N-1} r_i \alpha^{ki} \qquad 1 \leqslant k \leqslant d-1 \qquad (18)$$

is an evaluation of a polynomial of length N on $d-1$ points. Since $N > d-1$, it is best to compute all syndromes simultaneously in the following manner as each r_i is received:

$$S_k = \left(\ldots (r_{N-1} \alpha^k + r_{N-2}) \alpha^k + \ldots + r_1 \right) \alpha^k + r_0 \qquad (19)$$

Note that r_{N-1} is the first received symbol. Starting from the innermost parentheses, syndrome S_k is gradually computed as r_i are received. After r_0 is entered, all $d-1$ syndrome computations are completed at the same time. They are ready to be shifted out serially at that point. A systolic array design of a syndrome computation circuit is shown in Fig. 2.

IV. A VLSI Design for Polynomial Expansion

Recall that Λ is the set of α^{-i} where $\alpha^{-i} \epsilon \Lambda$ implies the location of r_i is an erasure. The computation of the erasure locator polynomial $\Lambda(X)$ demands the expansion of

$$\Lambda(X) = \prod_{\alpha^{-i} \epsilon \Lambda} (X - \alpha^{-i}) \qquad (20)$$

from one root α^{-i} at a time. Similarly, the modified syndromes

$$T(X) \equiv S(X)\, \Lambda(X) \bmod X^{d-1}$$

$$\equiv S(X) \prod_{\alpha^{-i} \epsilon \Lambda} (X - \alpha^{-i}) \bmod X^{d-1} \qquad (21)$$

can also be computed in the same manner except $T(X)$ uses $S(X)$, instead of 1, as an initial condition. Therefore, a polynomial expansion circuit is developed to calculate $T(X)$ and $\Lambda(X)$.

Note that for an arbitrary $S(X)$, which may be 1,

$$S(X)(X - \alpha^{-i}) = XS(X) - \alpha^{-i} S(X) \qquad (22)$$

This computation can be accomplished by a linear shift of $S(X)$, multiplication of every coefficient of $S(X)$ by α^{-i}, and finite field additions. A systolic array is designed, as shown in Fig. 3, to implement such simple operations. The control signal "zero" ensures that the resultant polynomial would not be changed if $\alpha^{-i} = 0$.

V. A New Architecture to Perform the Modified Euclidean Algorithm

A systolic array was designed in [2] to compute the error locator polynomial by a modified Euclidean algorithm. The array required $2t$ cells, twice the number of correctable errors. It is capable of performing the modified Euclidean algorithm continuously.

In the modified Euclidean algorithm only one syndrome polynomial is computed in the time interval of one code word. As a consequence, the original architecture in [2] of a pipeline RS decoder is not as efficient as it might be. A substantial portion of the systolic array is always idling. This fact makes possible a more efficient design with fewer cells and no loss in the throughput rate.

For the (N, I) RS code the length of the syndrome polynomial is $N - I$. The maximum length of the resultant Forney syndrome polynomial is also $N - I$. Imagine now that a single cell is used recursively to perform the successive steps of the modified Euclidean algorithm instead of pipelining data to

the next cell. Then it would take $N - I$ recursions to complete the algorithm, where each recursion requires $N - I$ symbol times. Therefore, using a single cell recursively requires only a total of $(N - I)^2$ symbol time to complete the modified form of Euclidean algorithm. Since a syndrome polynomial needs to arrive every N symbol times, only $\lfloor (N - I)^2/N \rfloor$ cells are needed to process successive syndrome polynomials at a full pipeline throughput rate.

Figure 4 shows the new alternate architectural design. The input multiplexer directs the syndrome polynomials to different cells. Each processor cell is almost identical to the cell presented in [2], except that it is used to process data recursively.

The architecture of the new basic cell is given in Fig. 5. Compared with the previous systolic array design [2], the present scheme for multiplexing the recursive cell computations significantly reduces the number of cells and as a consequence the number of circuits. Table 1 shows that the cell reduction is greater for high rate codes.

VI. A VLSI Design of a Polynomial Evaluation Circuit

In RS decoding the errata locator polynomial

$$\sigma(X) = \sum_{i=0}^{e+E} \sigma_i X^i \tag{23}$$

its derivative

$$\sigma'(X) = \sum_{\substack{i=0 \\ 2 \nmid i}}^{e+E} \sigma_i X^{i-1} \tag{24}$$

and the errata evaluator polynomial

$$\omega(X) = \sum_{i=0}^{e+E-1} \omega_i X^i \tag{25}$$

all need to be evaluated for each α^{-i}, $1 \leqslant i \leqslant N$. Note that the syndrome computation is another form of evaluating the received message polynomial $R(X)$:

$$S_k = R(X)\Big|_{X=\alpha^k} = \sum_{i=0}^{N-1} r_i X^i \Big|_{X=\alpha^k} \quad \text{for } 1 \leqslant k \leqslant d-1 \tag{26}$$

However, the syndrome computation is an evaluation of a polynomial of length N on $d - 1$ points and both $\sigma(X)$ and $\omega(X)$, having length $\leqslant e + E + 1 \leqslant d - 1$, are evaluated on N points. If one evaluates $\sigma(X)$ or $\omega(X)$ using the design in Section III for syndrome computation, it would take N of these cells. Since $N > d - 1$, there is a more efficient design which uses only $d - 1$ cells with less complexity.

Consider evaluating a polynomial $A(X)$, $\deg(A(X)) \leqslant d - 2$

$$A(X) = \sum_{i=0}^{d-2} a_i X^i \tag{27}$$

for $X = \alpha^{-j}$, $j = 1, 2, \ldots, N$.

Hence,

$$A(X)\Big|_{X=\alpha^{-j}} = \sum_{i=0}^{d-2} a_i \alpha^{-ji} = \sum_{i=0}^{d-2} a_i \alpha^{-ij} \quad \text{for } j = 1, 2, \ldots, N \tag{28}$$

For each a_i, the quantity $a_i(\alpha^{-i})^j$ can be obtained by recursively multiplying a fixed constant α^i as j goes from 1 to N.

A finite field summation of $d - 1$ terms results in the desired polynomial evaluation. A systolic array design of such an operation is shown in Fig. 6. Note that the results of evaluating $\sigma(X)$, $\sigma'(X)$, and $\omega(X)$ are produced sequentially. This matches perfectly with the sequential nature of the received data $R(X)$ in a real-time decoding environment.

One last observation on the polynomial evaluation: the evaluation of $\sigma'(X)$ uses only the coefficients of $\sigma(X)$ with odd power terms. This property makes it possible to obtain the evaluation of $\sigma'(X)$ as a by-product from the evaluation of $\sigma(X)$ at no cost. As illustrated in Fig. 7, simply use two smaller exclusive-OR trees to sum the even terms and odd terms of $\sigma(X)$ separately. The summation of the odd terms yields $\sigma'(\alpha^{-i})$. Another exclusive-OR operation on the two partial sums results in $\sigma(\alpha^{-i})$ itself.

References

[1] H. M. Shao, T. K. Truong, L. J. Deutsch, J. H. Yuen, and I. S. Reed, "A VLSI Design of a Pipeline Reed-Solomon Decoder," *IEEE Trans. on Computers*, vol. C-34, no. 5, pp. 393–403, May 1985.

[2] R. P. Brent and H. T. Kung, "Systolic VLSI Arrays for Polynomial GCD Computations," Dept. Computer Science, Carnegie-Mellon Univ., Pittsburgh, Pennsylvania, 1982.

[3] H. M. Shao, T. K. Truong, I. S. Hsu, L. J. Deutsch, and I. S. Reed, "A Single Chip VLSI Reed-Solomon Decoder," *TDA Progress Report 42-84*, October–December 1985, Jet Propulsion Laboratory, Pasadena, California, pp. 73–81, February 15, 1986.

[4] I. S. Reed, T. K. Truong, and R. L. Miller, "Decoding of BCH and RS Codes with Errors and Erasures Using Continued Fractions," *Electronic Letters*, vol. 15, no. 17, pp. 542–544, July 1979.

[5] T. K. Truong, I. S. Hsu, I. S. Reed, and W. L. Eastman, "Simplified Procedure for Correcting Both Errors and Erasures of a Reed–Solomon Code Using the Euclidean Algorithm," *TDA Progress Report 42-91*, July–September 1987, Jet Propulsion Laboratory, Pasadena, California, November 15, 1987.

[6] T. K. Truong, I. S. Hsu, I. S. Reed, L. J. Deutsch, E. Satorius, and H. Shao, "A Comparison of the VLSI Architecture for Time and Transform Domain Decoding of Reed-Solomon Codes," to appear in *Proc. ICCD '87*, Port Chester, New York, October 5-8, 1987.

[7] R. J. McEliece, *The Theory of Information and Coding*, Reading, Massachusetts: Addison-Wesley Publishing Company, 1977.

Table 1. Comparison of the number of cells required in the modified Euclid's algorithm computation

RS code	Full systolic array	Multiplexing on recursive cells
(15, 9)	6	3
(31, 15)	16	9
(255, 223)	32	5

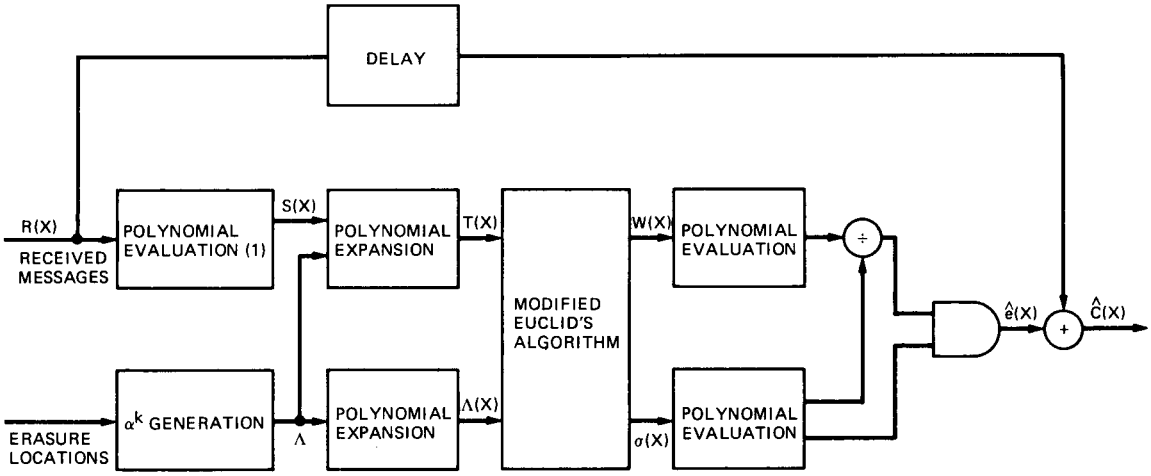


Fig. 1. VLSI architecture of a pipeline time-domain Reed-Solomon decoder for both error and erasure correction

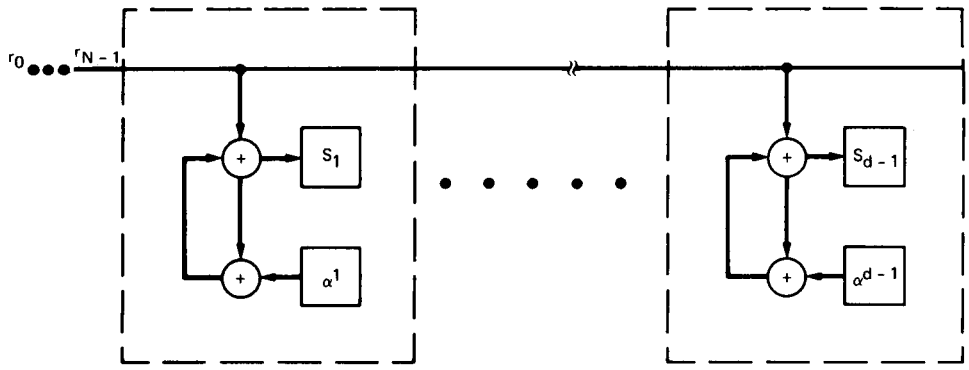


Fig. 2. A systolic array to compute syndromes

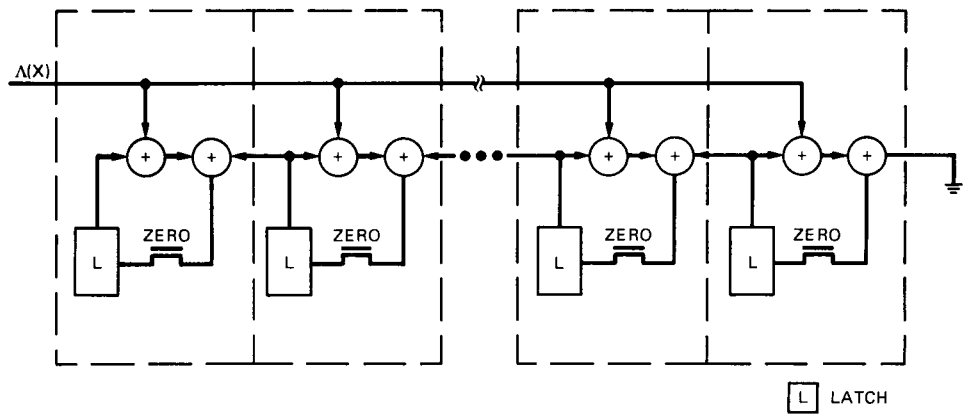


Fig. 3. A systolic array for polynomial expansion computation

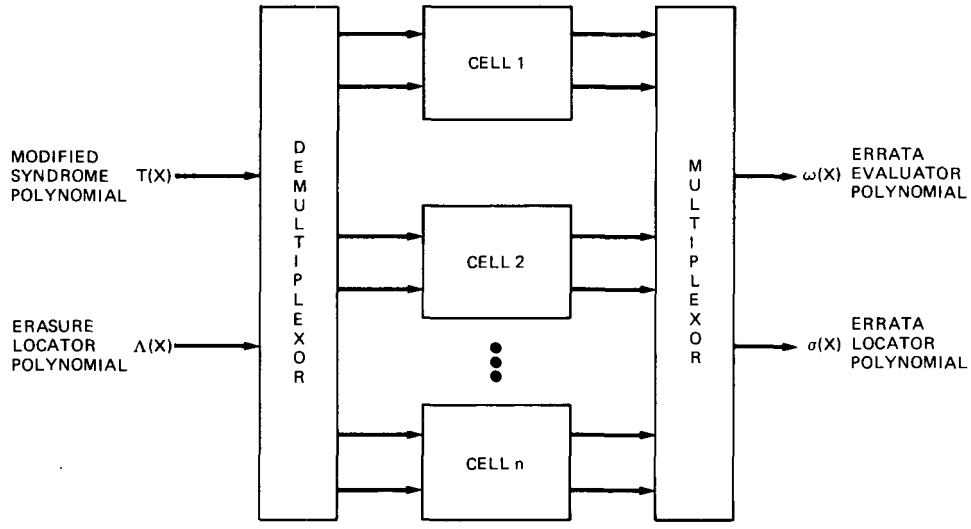


Fig. 4. The new architecture for performing the modified form of Euclid's algorithm

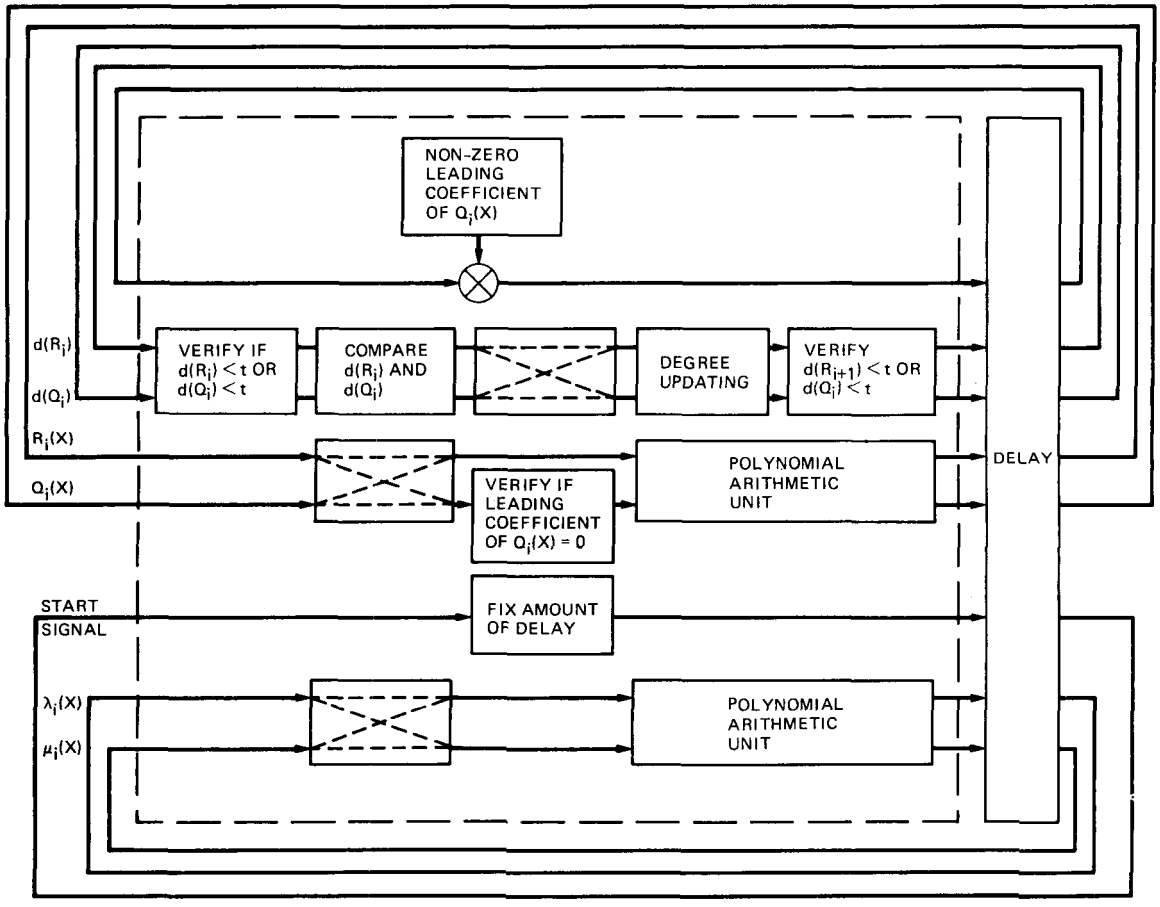


Fig. 5. Block diagram of basic cell for computing the modified Euclid's algorithm

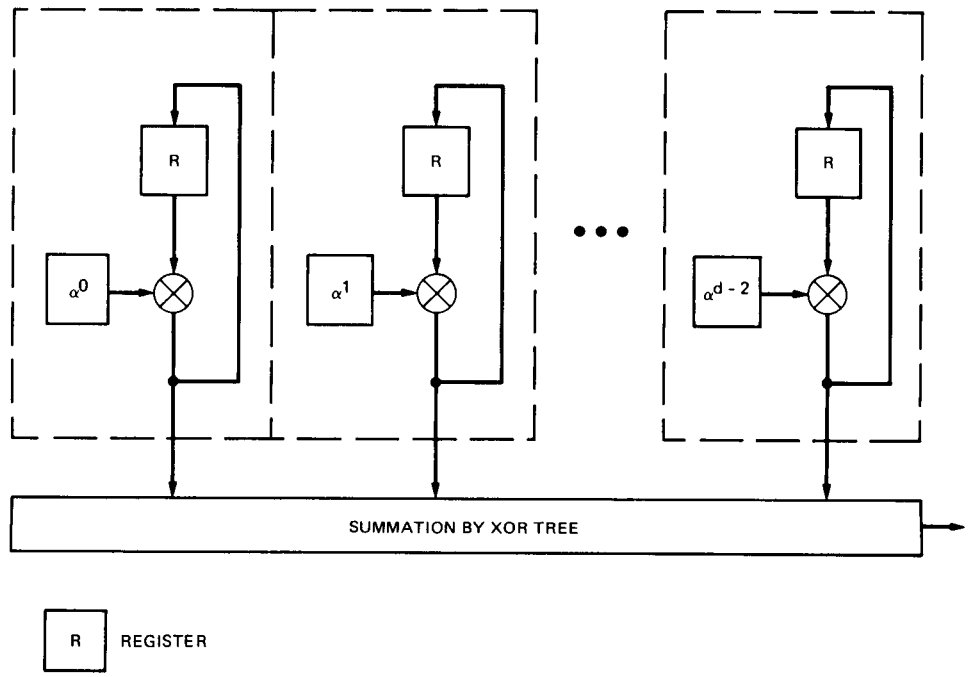


Fig. 6. A systolic array for polynomial evaluation

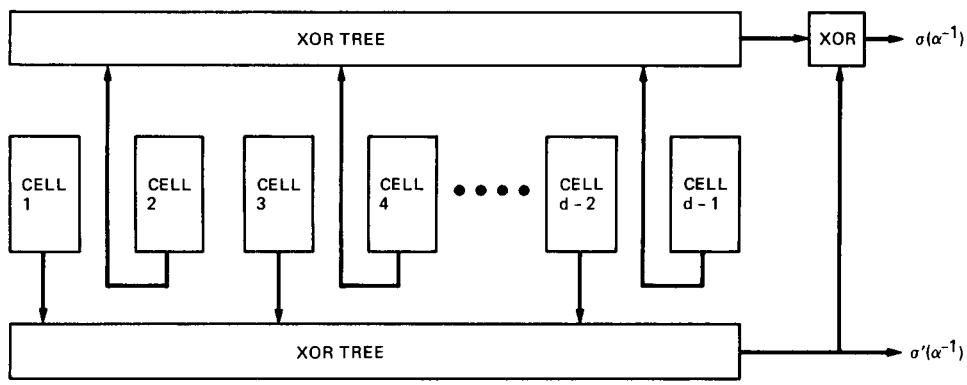


Fig. 7. The polynomial evaluation circuit for $\sigma(X)$ and $\sigma'(X)$

A Procedural Method for the Efficient Implementation of Full-Custom VLSI Designs

P. Belk
Flight Computer Systems and Technology Section

N. Hickey
NDH Consulting

An imbedded language system for the layout of VLSI circuits is examined. It is shown that through the judicious use of this system, a large variety of circuits can be designed with circuit density and performance comparable to traditional full-custom design methods, but with design costs more comparable to semi-custom design methods. The high performance of this methodology is attributable to the flexibility of procedural descriptions of VLSI layouts and to a number of automatic and semi-automatic tools within the system.

I. Introduction

Traditionally, full-custom integrated circuit design has been used in those situations where high performance or small size was of critical necessity. The designer has been forced to accept the increased development time and susceptibility to error over the so-called semi-custom approaches. This article is the first in a series of articles which describe a full-custom design methodology that supports many of the structured concepts inherent in the semi-custom approaches, allowing most of the benefits of a full-custom design to be realized while avoiding the penalties normally associated with full-custom design.

In a full-custom design, the designer must specify in great detail the actual cell geometries, cell placement, and cell routing of the chip. In exchange, the designer is able to control the system performance parameters of speed, power consumption, and size. In a semi-custom approach, the designer must select from predefined cell geometries, and the cell placement and routing are constrained by the automatic placement and routing procedures associated with the system. In many designs, there is only a small portion of the chip which is critical to the performance. In others, there are a few unusual functions which are not normally found in standard systems. In still others, the functions exist but are too large or slow or limited to be used in design. Hence, what is needed is a system which will provide the automatic capabilities of the semi-custom approach but which will also provide easy access to the special leaf cells, special placement, and special routing procedures which are needed in a given design.

In addition to the functional differences between this approach and the traditional methods, there is a philosophical one. The apparent goal of many traditional systems is to eliminate the designer from the design cycle. Under the approach described in this article, this is not necessarily considered a

worthwhile goal. Rather, the goal of the system is to maximize the efficiency of the designer by allowing the utilization of automatic tools where appropriate, procedural tools where appropriate, and manual tools where appropriate and to make it easy to determine when each is needed. Furthermore, a single, well-defined interface between all the tools allows the designer to use his intuition and intellect on those design aspects requiring the greatest design power. The interface to the system is based on the premise that the intuition and skill of the designer are valuable resources.

The central element of the layout system described is an imbedded language called Art. Art is compatible with a set of other tools created as part of the system, including editors, plotting programs, and design rule checkers. The methodology described in this article presumes that most of the chip layout task will be accomplished by using Art to perform placement and routing between pre-existing low-level cells. Although this is a very efficient method of implementing a large variety of very complex chip designs, it should be emphasized that there is no element in this system which mandates the use of any feature of an imbedded language for any part of the design. Thus, the designer may choose to use only Art's predefined placement and routing capabilities, to define more advanced or specialized placement and/or routing capabilities in Art, or to bypass Art entirely in favor of implementing the entire design in the graphics editor or other layout tools.

The low-level cells used by Art may be created by other components in the system, or imported. Imported cells may come from existing standard cell libraries, external cell generators, or other types of external design systems. Those generated internally may be manually generated (in the interactive graphics editor) or procedurally generated at any level. Procedural generation of cells is normally limited to composite cells (cells created by combining previously defined cells) and generated cells (such as PLAs and ROMs), but can optionally be used for the creation of special types of cells, such as technology-independent cells, or cells having configurable speed or drive capabilities. Consequently, the system provides for the combination, in a single design, of cells from a variety of different sources, and the designer may determine which sources best satisfy the requirements of the design.

The important design information is stored in a small collection of design databases by Art. Most of this information is contained in a cell database for each cell in the design. When cell information is read or created by the system, it is distilled down to a basic external description of the cell which contains only that information necessary to utilize the cell as a component in other cells. Thus, the size and amount of time necessary to scan the design database are minimized. This feature offers the flexibility to create design methods which process the entire external cell interface while still allowing a relatively short design loop.

Placement and routing of cells within Art is facilitated by a number of automatic routines. These include the "Tiler" program, which produces data-path-like structures, and an automatic channel router, "MidBus," which implements a very general two-layer routing algorithm for random interconnect between rows or columns of cells. Where these automatic tools are inappropriate, the designer may specify the placement and routing through easily written procedures. The designer may utilize any or all of these methods in a given design.

All of the tools are written in and compatible with Art, and thus the designer is able to choose the appropriate tool independently for each part of a layout. Hence, the designer is always able to be as specific as he desires in controlling the action of the system. If none of the existing tools is appropriate to the task at hand, he is free to modify an existing tool, to create an entirely new special-purpose tool, or even to directly specify the design procedurally. The designer is thus not forced to distort a design in an attempt to map an otherwise incompatible design to the capabilities of the system.

This methodology is sufficiently flexible to allow a large number of unique problems to be addressed. The example chip in this article required the generation of regular and irregular rectangular arrays of cells, the hierarchical interconnection of functional blocks, the use of a PLA, and the layout and interconnection of random logic. The designers were able to use the data-path compiler, the channel router, the automatic PLA generator, and various procedural place and route functions to complete the chip. It will be shown that the combination of these techniques provided an efficient, well-structured, and easily maintained design.

Other articles will deal in more detail with the generation of specialized cell generators, including an in-depth examination of the PLA generator and an examination of methods for producing technology-independent generated cells. The use of Art for the efficient generation of procedural place and route, such as its use in the control block, will be described. Also, the interface between this layout system and external simulation and verification tools will be discussed.

II. History of the System

The concept of using an imbedded language approach was first developed at the California Institute of Technology in 1977 with the creation of LAP [1]. As used at Caltech, however, this tool was regarded primarily as a method for creating low-level cells rather than for the assembly of complex higher-level cells [2]. Subsequent work of the Caltech Silicon Struc-

tures Project (SSP) proceeded along a path toward a "silicon compiler." The design system would require the designer to specify connectivity information and, possibly, relative cell placement information, from which the geometry would be created in an automated manner [2]. Examples of this approach include a system for the implementation of data paths, the Bristle-blocks system written by D. Johannansen in 1978 [3], and Earl [2].

The original Art program was developed at JPL in 1980 by John Wawrzynek using many of the concepts present in LAP. This version of Art was written in Pascal and supported only the capability of creating the basic primitive geometrical elements found in Caltech Intermediate Form (CIF) [4]. This original Art suffered from the requirement that a completely new version was needed for each new CMOS or NMOS technology.

In 1982, a completely new version of Art was produced at JPL by Paul Belk and Steve Trimberger. This Art supported some of the concepts of the current system, specifically the use of symbol names, the definition of connection points, and the maintenance of a design database for each cell. This system evolved over the next few years to include a number of semi-automatic routers and cell generators but still suffered from the requirement of a separate version for each technology.

In 1986, Art was completely rewritten under the C language. This version of Art was designed to be technology-independent and to support the generation of technology-independent designs. In the course of producing this version, a number of more sophisticated placement and routing tools were developed. In addition, the dependence on CIF as the geometric database was eliminated; the parser and generator for the geometry files were separated from the other code to allow the substitution of other geometrical formats. Furthermore, the concept of logical names for symbols and ports was extended to include instance names, and the ease of access and completeness of each cell's database were substantially improved.

Since its initial introduction, Art has seen substantial improvements and additions. It is expected that this enhancement will continue because of the ease with which these additions can be made to the basic system.

III. Description of Design

The sample chip to be described in this article is part of a multi-processor signal processing unit. The chip is responsible for controlling access to dual-port RAM, providing sequential RAM access for one of the CPUs, and providing a collection of utility counters and timers. In addition, it supports an autonomous data collection mode during the time that both CPUs are powered down.

The chip was chosen for this article because it provides an effective demonstration of the combination of several disparate design techniques on a single chip. At the highest level, the chip contains a data-path-like structure containing a large number of standard utilities, such as counters, latches, shift registers, and buffers; and a control section containing highly regular structures (including a PLA), irregular structures (clock generators), and completely random logic. It will be shown that, with the exception of lowest-level leaf cells, each containing about 10 simple gates and numbering less than 20, it was possible to efficiently implement all elements of the chip in Art using automatic or semi-automatic methods.

The chip is divided into four functional units (Fig. 1). These units are discussed below.

A. High-Speed Block

The high-speed block contains the cells which control access to the RAM by the two CPUs. Specifically:

(1) a 3-word by 16-bit buffered read FIFO;

(2) a 3-word by 16-bit buffered write FIFO;

(3) a 16-bit starting address latch;

(4) a 16-bit auto-increment address pointer; and

(5) a 16-bit bidirectional tristate buffer between the high-speed block and the low-speed block.

B. Low-Speed Block

The low-speed block contains a number of utility blocks, including timers, address latches and pointers, serial-to-parallel and parallel-to-serial converters, and command state latches. Specifically:

(1) a 16-bit address multiplexer;

(2) two 16-bit latches;

(3) a 16-bit auto-increment address pointer;

(4) 16-bit serial-to-parallel and parallel-to-serial converters;

(5) 6 4-bit to 25-bit timers;

(6) a 5-bit, double-buffered command latch; and

(7) a 16-bit test shift register.

C. Control Block

The control block contains the PLA which controls all other devices on the chip. It also contains circuits for condi-

tioning off-chip inputs and outputs and for producing the two-phase clocks required by other circuits. Specifically:

(1) a 23-word by 42-bit control PLA;

(2) an inverting Schmitt trigger for reset conditioning;

(3) two one-shot circuits for edge detection;

(4) 7 flip-flops for conditioning of input signals;

(5) 3 two-phase, non-overlapping clock generators for producing chip clocks; and

(6) synchronization logic for synchronizing memory requests from the low-speed block with the high-speed clocks.

D. RAM Access Arbiter Block

The RAM access arbiter block contains the random logic necessary to control RAM access between the two CPUs and the chip. It also creates the signals that increment the RAM address pointers.

IV. Description of Methodology

Much of the power and generality of the approach described comes from the fact that the designer may solve a complicated design problem by breaking the problem into a number of separate stages, each of which utilizes a small number of simple and well-understood design techniques. The design is thus reduced to the successive application of these techniques. Since these techniques may be combined in a very flexible manner at each level, it is possible to implement a chip architecture which is quite complicated when viewed in its entirety, by successively breaking the design into architectural blocks, each of which in turn consists of a small number of blocks having a single simple, well-defined interface. Further, the implementation of each technique may be broken into a small number of simple primitives.

Since the methodology described in this article is a layout methodology, no consideration is given to other systems issues. In particular, it must be assumed that the system design has been properly partitioned to allocate reasonable functionality to the chip being implemented and that the system design as a whole has been simulated to verify that the chip, once produced, will operate properly within the system. It will also be necessary at the conclusion of the layout effort to provide for functional layout verification in the form of a layout vs. schematic check or a net list extraction. Layout design rule checking is also important but is usually performed continuously during the design. In the case of the chip described in this article, design rule checking was performed with the internal design rule checker DRC, which is compatible with both Art and the graphics editor.

As with any layout methodology, it is very important to devote sufficient design time at the beginning of the project to the development of an optimal design partitioning scheme. It is extremely important that, to the greatest extent possible, each stage of the design be reduced to a small number (normally less than 10) of self-contained blocks having a simple and well-defined interface. Although it is possible to implement a badly partitioned design using the methodology described, the design will almost certainly be significantly more prone to error and take much longer to implement.

The following sections provide a brief description of the primitives and techniques which were combined to implement the chip discussed. These concepts are a subset of those supported by this methodology but are sufficient to demonstrate its use.

A. Imbedded Languages

The power and flexibility of the design methodology described in this article are attributable to the use of direct procedural methods for specifying the placement and routing of objects based on logical references to size, location, and type attributes associated with those objects. The generation of the layout is thus reduced to a programming task utilizing all the power and flexibility that such an approach implies.

Art consists of the definition of various database structures containing information about the design, a subroutine library for creating and manipulating these structures, routines for converting these structures into the final geometrical information, and a collection of macros to allow easier access to the program functions. Instead of attempting to define a special purpose syntax for the specification of the design, Art is written in and accessed by programs written in C. Thus the designer is able to use all the functionality and power of a standard programming language. Simply stated, Art is a layout system imbedded in C, or more simply an "Imbedded Layout Language."

No assumptions about the nature of the geometrical design rules or the target fabrication technology need be made in Art. Art is very largely technology-independent in the sense that such details as design metric, mask layers, geometrical design rules, and the format of the final layout database are not built into Art. The design metric may be chosen by the designer, since the numbers used in Art may be taken to represent whatever measurement units are necessary. The names and natures of the mask layers are read by Art from a file unique to that technology, as are a set of geometrical design rules, and the layout database is read and written by separate input and output modules and converted to an internal database which is

used by Art. Art tends to assume a certain flexibility in terms of symbol hierarchy, but the system includes filters which may be used to flatten the hierarchy to whatever extent is necessary for the target mask pattern-generation technology. The chip described was produced for a standard 3.0-micron CMOS/Bulk technology and the mask data was transferred to the fabrication house in the Caltech Intermediate Form (CIF).

Art is not limited to use by experienced programmers. Although a skilled programmer may access the primitive functions directly to produce a layout, Art may also be used to generate various placement and/or routing functions which are then accessible to non-programmers. In the sample chip, both approaches were used. The data-path placement program Tiler, the channel router MidBus, and the PLA generator were written in Art by skilled programmers to allow simple non-procedural specification for major portions of the design. These programs may be used by non-programmers to solve a more general set of problems. In the control block, however, much of the placement and routing was performed using a few special purpose subroutines which directly accessed the Art primitive functions.

The capability of defining both general-purpose and special-purpose placement and routing functions within the same system allows for maximum efficiency in the design process. Furthermore, it is often possible to enhance a special-purpose solution to the point where it provides a more effective general purpose solution than the more traditional semi-custom approaches.

B. Symbols and Instances

At a basic level, a symbol is a portion of the design which has been grouped into a single self-contained logical unit by the designer for reasons of convenience. It is defined solely by its geometrical representation if generated external to the system, or by the imbedded language source from which the geometry is eventually created. The geometrical representation used in this system is extended to include abstract attributes including each symbol's name, size, connection points, and alignment points. When the symbol is processed by the system, a "symbol reference database" containing this abstract information is created. This database is associated with a program identifier identical to the symbol name.

When a (child) symbol is used within the definition of a (parent) symbol, only the abstract information contained in the child symbol's reference database is accessible. Thus, within the context of the parent symbol, the child's abstract attributes comprise the complete description of the child symbol.

This use of the child symbol is referred to as an instance. The instance database includes the reference to the child symbol, the physical location of the instance within the parent, and an optional instance name. If an instance name is provided, the location of the instance's connection and alignment points may be easily accessed. If no instance name is provided, this information is not readily available. The instance name, if provided, will appear in the extended geometrical representation of the parent symbol.

C. Ports: Connection and Alignment Points

When combining child symbols within the parent symbol, it is necessary to determine both the correct location for the child (relative to other items within the parent) and the locations on the child to which connections are to be made either by wire connection or by direct cell abutment. These locations are referred to as the child symbol's (or instance's) ports. Each port is described in the child symbol's reference database. This description includes the port's location on the child symbol, its name, its connection layer, and its type. For connection points, the connection layer indicates the material (e.g., metal or poly) of the wire which should connect to it; the type indicates the purpose (e.g., input, output, VDD, etc.) of the connection. If the port is used only for alignment purposes, then the material and type are left blank. All references to a port within the language are made using the instance name and port name.

In order to simplify the action of many of the special purpose interconnect procedures, the designer will often impose a naming standard for the ports. For example, it is common to define the upper right VDD port as "VDD_ur" and the lower left ground port as "GND_11" and to use these as the alignment points. Similarly, on signals which feed through the symbol vertically, the top port is referred to as "name_t" and the bottom port as "name_b."

D. Leaf Cells and Composite Cells

Previously, symbols were defined as a portion of the design grouped into a unit by the designer for convenience. It is useful to distinguish between several categories of symbols based on the symbol's use and internal structure. The simplest type of symbol, containing at most a few geometrical objects and no ports, which has been defined solely to allow easy access to a standard feature of the design technology, is referred to as a *macro symbol*. These symbols serve the same purpose as macros in a programming language. Common examples, used in most designs, are the inter-layer contacts.

More complex symbols, which provide a basic circuit function, are referred to as leaf cells. Leaf cells are the basic atomic building blocks of a structured design. Often these cells are

quite simple internally, consisting of only a small number of gates, but occasionally may be very large atomic structures such as PLAs, RAMs, or ROMs, which, though large, still cannot be readily described in terms of simpler functional units. When the cells are small and simple in function, they are frequently created by hand in an interactive graphics editor. When more complex, it is often useful to create special generators such as PLA or ROM generators to create the cell. Either type of leaf cell may also be easily imported from an external system which provides a library of cells or of cell generators.

Leaf cells approximate the function of standard cells in a standard cell system but differ from standard cells in two important respects. First, they are defined by the designer only as needed for the particular application; and second, they may be optimized, or customized, for the specific environment in which they will be used. Though this customization is not necessary for the design methodology to work, it provides a powerful means of increasing the efficiency of the final design.

Leaf cells are normally viewed as "black boxes" for both layout and function. Hence, in addition to the geometrical information, it is necessary to provide the abstract attributes of each leaf cell. Various methods have been provided for doing this. The graphical editor supports the direct entry of this information. When standard cells are used, it is often sufficient to convert the "footprint" data which is provided with the cell library.

As was described previously, the methodology is based on the synthesis of the final design by combining many symbols into more complicated symbols. These symbols, which are readily considered to be a collection of functional sub-blocks, are referred to as composite cells. The distinction between leaf and composite cells is extremely useful in understanding the application of the design methodology within the context of the generation of these symbols, but when a symbol is used as a child symbol for the generation of a parent symbol, this distinction becomes unimportant. That is, within this design methodology, leaf cells and composite cells may be used interchangeably as child symbols in the generation of higher-level parent symbols.

It is important to realize that a normal design will utilize at least 3 separate levels of source files. The results of each level are then used as if they were leaf cells for input to the next level.

V. Implementation Details

Within the Art methodology, a chip is designed by starting with a top-level architectural view and dividing the chip into a small number of relatively self-contained components having a well-defined interface. These components are then analyzed in a recursive manner to reduce them to simpler components in a similar manner. This process continues until the designer is able to identify an appropriate set of leaf cells with which the design may be implemented.

Once a preliminary leaf cell set has been identified, the designer may modify the chip design in order to minimize the number of leaf cells which must be created for the design. Also, he may determine which, if any, of the leaf cells are already available from external sources. He will also make a preliminary decision on the method to use to generate the leaf cells.

Having defined the design hierarchy and the leaf cell set, the designer must then resynthesize the complete chip. He does this by choosing the appropriate placement and routing algorithms with which to create each component. The designer must also consider the overall chip floor plan when implementing each component so that its size and port locations are well matched to the overall structure. It is not uncommon for a designer to modify the initial split-up of the design at this point in order to simplify the final design. Hence the split-up and synthesis should be viewed as an iterative process.

For the sample chip described in Section III, the top level was divided into components which corresponded to the four functional blocks. Each block required connections to the chip's pads and each block shared many control signals with every other block. Analysis of the interblock connections provided the following:

(1) The high speed and low speed blocks' data busses interconnect;

(2) The low speed block and control block have many common signals;

(3) The high speed and arbiter blocks have many common signals; and

(4) The high speed block has most of the data bus pad connections.

Thus, the most practical layout had the high speed block above the low speed block on the left of a central routing area and the arbiter block above the control block to the right. The data busses for the high speed block were routed from the middle of the high speed block to pins on the left; and across the chip, between the control and arbiter blocks, to additional pads on the right. Details of the layout methods used for the top level of the chip are described in the following section.

A. Top-Level Layout

Once the top-level functional division and floor plan were decided on and each of the four functional blocks generated, it was possible to produce the actual top-level geometry. The generation of this geometry used a number of the techniques which are supported under the Art system. The sequence of operations was as follows:

(1) Initialize the Art system and open output files;

(2) Read in geometry information for the sub-blocks;

(3) Scan each sub-block to obtain a list of all its ports;

(4) Place high speed and low speed (PMC) blocks at final locations, calculate y positions of arbiter and control blocks (x position resolved later);

(5) Generate the signal list and tie points for the central bus;

(6) Allocate bus channels and determine bus width;

(7) Place arbiter and control blocks;

(8) Draw the main power grid;

(9) Draw the middle control bus;

(10) Process the non-bus ports; and

(11) Close the symbol definition and the output file.

The main program which performs the above steps is a fairly compact three pages of code. The resulting layout is shown in Fig. 1.

B. Port List Scanning

One of the more tedious aspects of entering the layout program is the individual description of the types of interconnect that are required for each individual port in the subcells. In the top level of the sample circuit, it was determined that all of the ports on each cell fell into one of a few groups. The actions necessary for each port in the group were easily specified. Since one of the actions supported in the Art system is the ability to scan each port of a symbol, the layout program was able to generate the necessary two lists of port names for each symbol.

All ports which were to be connected to the central data bus were given names starting with the characters "bus_". These characters were followed by the name of the control signal to which it was to be connected plus an optional "_1", "_2", etc. For example, port "bus_phi2" would be connected to control line "phi2", and ports "bus_phi1_1" and "bus_phi1_2" would both be connected to control line "phi1". The control bus connectivity was thus readily available directly from the port names themselves.

Similarly, the VDD, GND, and pad I/O ports were provided with names which indicated the correct connective action. For example, the low speed block supported automatic checking of the interconnect with the high speed block (ports starting with "lp_"), pad connection ports on the bottom of the block ("pd_"), pad connections on the left ("pl_"), and VDD and GND connections. Any port whose name did not start with one of the expected prefixes would cause an error message to be displayed.

At first glance, restricting port names to start with the prefixes above might seem to place an unreasonable burden on the designer. In fact, however, all the port names were generated procedurally in the lower blocks and the final interconnect was verified with a layout vs. schematic tool. Hence, the top-level interconnection was produced with very little effort.

C. MidBus Router

The large numbers of control signals which were interconnected between the four sub-blocks of the chip suggested the use of an automatic channel router. Since no channel router existed in the system at the time this chip was implemented, it was decided to produce one specifically to satisfy the requirements for this chip.

The first issue resolved was the router type. Due to the preexisting power bus routing, it was decided to use a simple channel router utilizing a vertical metal channel with horizontal connections to the sub-block ports made in metal-2. Furthermore, it was decided to support the sharing of multiple signals within a given vertical channel but not to allow a signal to jog between vertical channels. This resulted in a very fast procedure which produced acceptable interconnect routing for this chip.

Using the port list which was generated as described in the previous section, it was a simple exercise to generate the list of signal names and tie points. An initial list of signals was also provided in the source code as a check for any missing signals. Finally, a list of the control signals which were to be connected to the chip's pads was also included.

The generation of the entire control bus was thus reduced to the following steps:

(1) Define control signals with external (i.e., pad) connections;

(2) Add an initial list of expected control signals (for extra check);

(3) Generate a complete list of signals and tie points from port scan;

(4) Allocate vertical channels to signals (and determine bus width);

(5) Place each block at the calculated location; and

(6) Create the geometry for the busses.

Note that the routines utilized in steps 4 and 6 constitute what is considered the MidBus router. It is thus able to produce a bus from any list of signal names and tie points.

D. Sub-Block Implementation

Having defined the interface between the top-level cells, it was necessary to determine the design technique most appropriate for each of the cells. The highly regular structure of the high and low speed blocks, along with the fact that each of these blocks operated on three common data busses and had a minimum of other interconnections (except for control signals), made them ideal candidates for layout using a datapath layout tool. Such tools are generally limited to bus-type structures but often prove to be the most efficient implementations of such structures.

The arbiter block was composed mainly of random logic and had very little internal regularity. Large amounts of random logic are notoriously vulnerable to layout error and often require many design iterations, but it was determined in this case that the required random logic could be reduced to a small number of leaf cells, each containing fewer than 10 simple gates. The cells were then placed in two columns and routed together using the MidBus router (Fig. 2).

The control block also lacked the regularity apparent in the high and low speed blocks but could make use of many of the lower-level primitive subroutines which had been developed as part of Tiler to allow the direct specification of its internal placements and wiring. The PLA in the control block required the generation of a special-purpose CMOS/Bulk PLA generator which accepted assembler-like input describing the state machine and generated a PLA cell. Neither of the special techniques used within the block is within the scope of this article, and the control block is thus best treated as a leaf cell within this context.

E. Tiler Datapath Compiler

Two major sub-blocks of the chip, the high speed and low speed blocks, are easily represented as rectangular arrays of leaf cells. It was determined early in the design cycle that the optimum arrangement is a vertical slice for each functional unit with 3 common data busses routed between each bit. Control logic for each unit is contained in a leaf cell at the top of each column, and the control lines run vertically through each cell.

Placement of the cells in each column is defined by a simple "COLUMN" macro. This macro creates a data structure containing the name of the active cell, the name of the control cell, the active bit positions, and whether the cells should be placed with or without first mirroring them. In addition, a number of data arrays are defined for each column. The "1_name" array provides a list of the ports which must line up vertically inside the column. This information is used by the system to verify the leaf cell design. The "x_name" arrays contain a list of contact points for the contacts to the busses and for those cells which are contact programmable. The "p_name" arrays provide the port assignments for the external control signals. Using the information provided in the structures just described, after placement of each column, the system would automatically define the external control ports, verify the control signal alignment, and place contacts at the appropriate locations for bus connections and configuration.

In addition to the three busses, adjacent columns may connect directly together. Placement of these input and output ports was performed in an *ad hoc* manner and then verified by the interconnect procedures. Figure 3 shows the low-speed block. It should be noted that most sections of the low-speed block have a very high density of active circuitry. Those sections which are empty are due to the fact that some of the elements in this block are less than 16 bits wide and therefore did not require a full bus slice. The high-speed block was composed almost entirely of 16-bit slices and thus is uniformly dense.

F. Probe Point Generator

In any complicated chip, it is useful to provide for direct probe access to verify the chip operation and/or determine the reason for non-operation of the prototype units. On the sample chip, it was determined that the majority of signals of interest were present on the middle control bus. In addition, it was determined that there would be ample space between the control block and the bus arbiter on the right side of the chip, directly adjacent to the middle control bus. This provided an opportunity to create two arrays of probe points, which would provide the capability to sample any signal on the control bus.

The creation of the probe point arrays was accomplished by writing a simple routine in Art which created a metal pad with a large enough cut in the passivation layer to allow the metal to be directly contacted by a normal chip probe. A wire was run from the center of the metal pad to the inside periphery of the probe array, and a port was placed at the end of the wire. To provide for the probing of each signal of interest, a list was

compiled of all signals in the control bus. That list was then scanned in a loop, and a probe point was created for each signal in the list. When this cell was added to the top-level chip, the MidBus router automatically scanned the ports at the periphery of the cell and connected each of the probe points to the appropriate signal on the control bus.

As can be seen in Fig. 4, the addition of the probe point array significantly changed the topography of the chip, especially the width of the MidBus, due to the increased length of most of the signal lines, which had previously been much more local. With the exception of the generation of the routine that created a single probe point in the array (less than one page of code), the entire process was automatic. The generation of the probe point array required a total of approximately 4 man-hours. Further, the difference in the characteristics of the control bus before and after the addition of the probe points provides a reasonable indication of the efficiency of MidBus in use of space.

VI. Summary

In this article, a methodology was presented, the Art methodology, in which an imbedded language is used to manage the complexity of a full-custom VLSI design. It was shown that the use of this methodology significantly reduced the complexity faced by a designer at any stage of a design without appreciably reducing the density or performance of that design. Thus, the Art methodology provides most of the benefits of traditional full-custom integrated-circuit design while substantially reducing the design costs.

Unlike many semi-custom design approaches, the Art methodology does not impose any inflexible constraints on the design itself. Several methods are shown which make the implementation of circuit structures less difficult, and through the application of combinations of these methods, a large variety of circuits may be implemented. Further, the designer is always free to disregard any of the existing methods in favor of developing special-purpose techniques for the implementation of specific structures. Although this development effort may increase the cost of the design, it is shown that many such techniques can be developed quite efficiently within the context of Art, and, more importantly, that such development will not affect the implementation of other parts of the design. Thus, the designer is allowed to make individual cost/benefit trade-offs on whichever portions of the design are considered critical.

Art provides considerable power beyond the management of high-level design complexity. Since Art is essentially technology-independent, it can be used for the development of designs in a wide variety of technologies, allowing rapid adaptation to the new circuit design technologies and techniques. In addition, the independence of the functional aspects of Art from the final representation of the mask geometry allows the adaptation of Art to diverse fabrication technologies and the implementation in Art of more fabrication-specific constructs, such as alignment marks.

The chip discussed in this article provides only one example of the power of the Art methodology. The chip was chosen because it demonstrated several different implementation techniques, but there are additional techniques that may be explored by examination of other types of chips. The chip described, however, provides useful information about the efficiency of the Art methodology. Because it was possible to use the same leaf cell in multiple contexts, it was possible to reduce the number of leaf cells designed to approximately 20. The layout of these cells required approximately 8 man-weeks. After the completion of leaf cell design, the implementation of the entire chip required three weeks for a two-person design team. Thus, a 14-man-week design effort resulted in the implementation of a chip having approximately 12,000 transistors. A small number of errors were detected in the chip by subsequent automatic layout vs. schematic comparison, ranging from errors in top-level interconnection to errors in leaf cell design. These errors required less than a day to correct.

Acknowledgment

The authors wish to acknowledge Tom Bulgerin, Linda Lee, Peter Jones, and Tim Shaw for their contribution to the development of the layout system and the chip used as an example in this article. The authors also gratefully acknowledge Dr. William Whitney for his insight and advice in the development of the topics discussed.

References

[1] C. R. Lang, *LAP User's Manual*, California Institute of Technology Computer Science Department Technical Report No. 3356, 1979.

[2] S. Trimberger, "A Structured Design Methodology and Associated Software Tools," *IEEE Transactions on Circuits and Systems*, vol. CAS-28, no. 7, July 1981.

[3] D. Johannansen, "Bristle Blocks—A Silicon Compiler," in *Proceedings of the 16th Design Automation Conference*, 1979.

[4] C. A. Mead and L. A. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

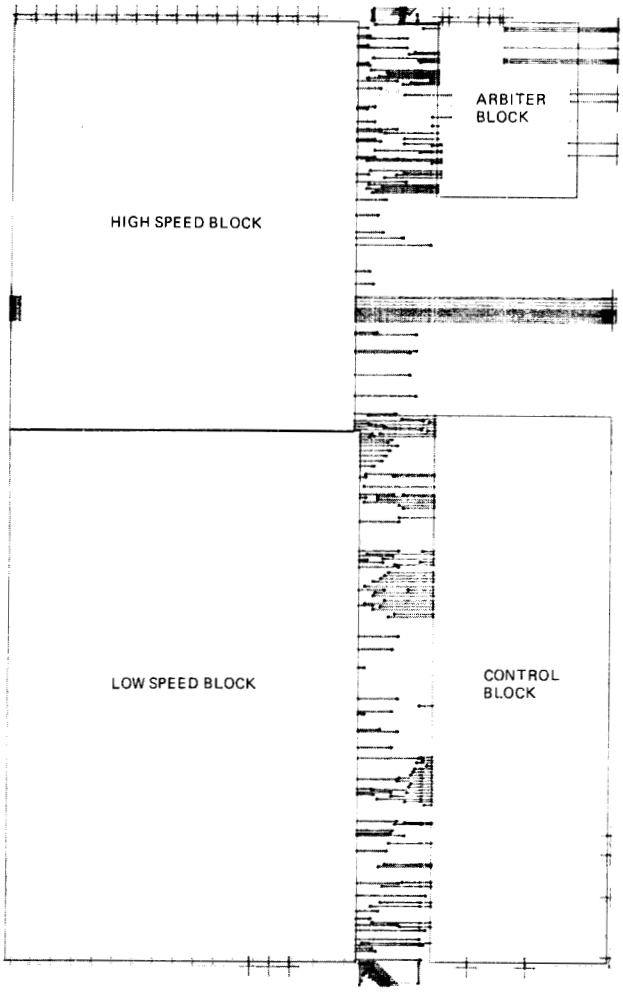


Fig. 1. Top level of sample chip

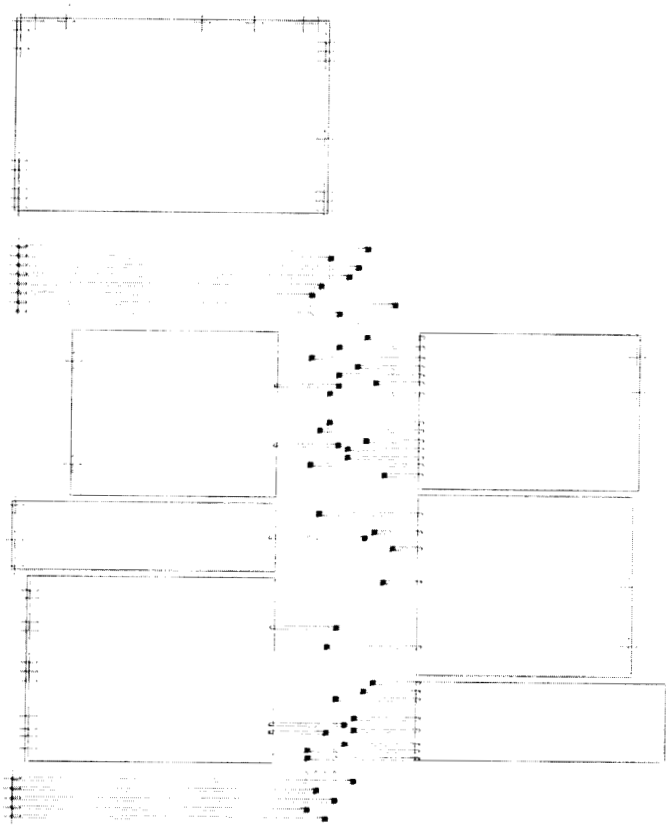


Fig. 2. Arbiter block

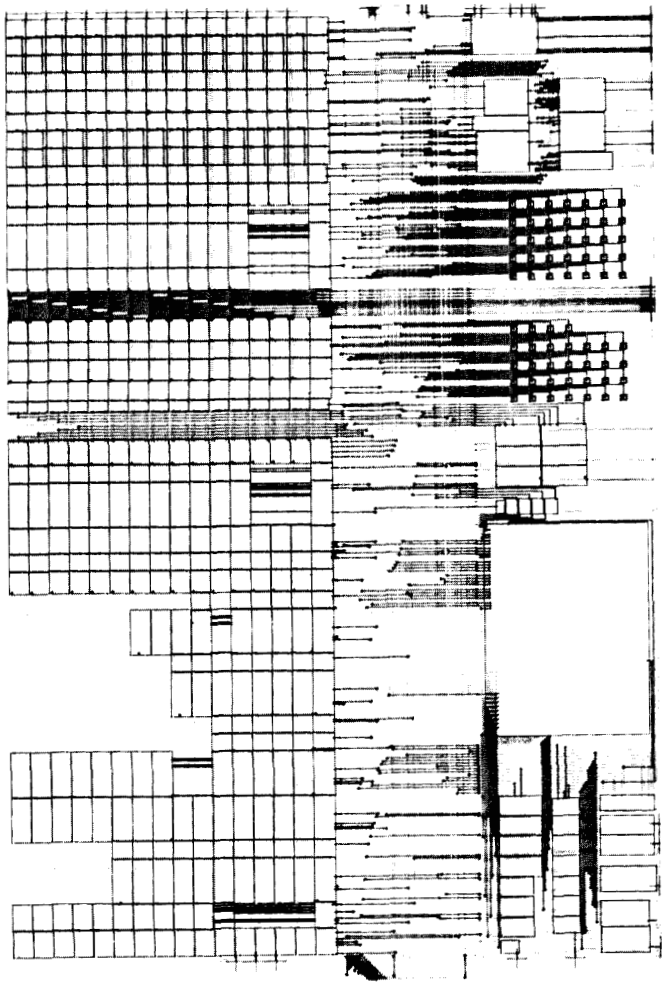


Fig. 3. Low speed block

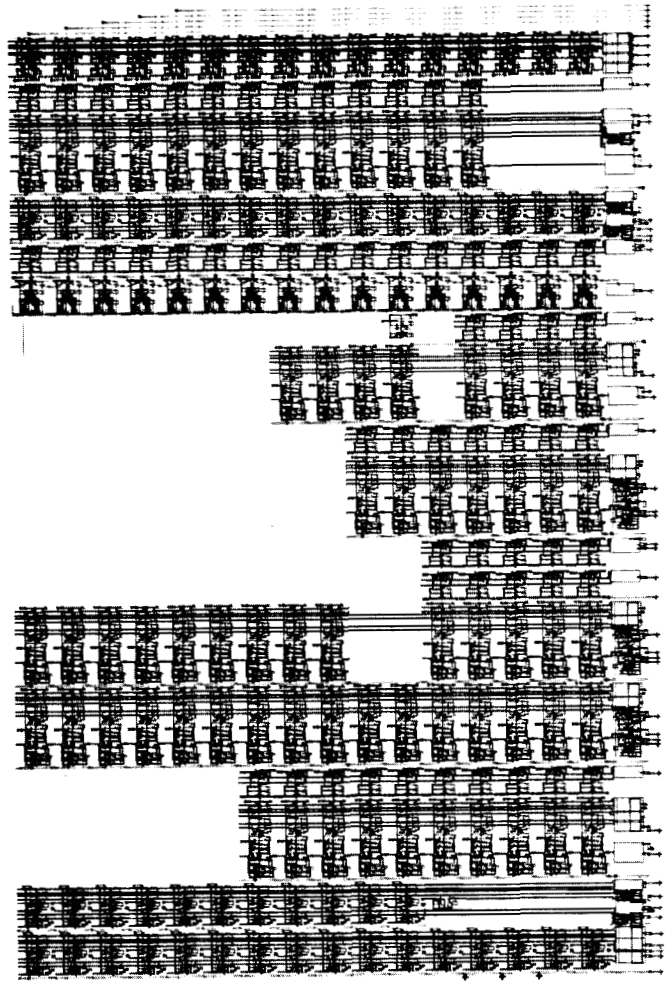


Fig. 4. Top level of sample chip with probe points

A New State Space Model for the NASA/JPL 70-Meter Antenna Servo Controls

R. E. Hill
Ground Antenna and Facilities Engineering Section

A control axis referenced model of the NASA/JPL 70-m antenna structure is combined with the dynamic equations of the servo components to produce a comprehensive state variable (matrix) model of the coupled system. An interactive Fortran program for generating the linear system model and computing its salient parameters is described. Results are produced in a state variable, block diagram, and in factored transfer function forms to facilitate design and analysis by classical as well as modern control methods.

I. Introduction

The upgrade of the NASA/JPL 64-m antennas to 70-m apertures has added considerable mass and inertia moment on top of the existing alidade with resultant decreases in the natural frequencies of the structure. Because the combined compliances of the alidade and gear reducers separate the autocollimator from the antenna servo motors and tachometers, the increased inertia can affect the dynamics of autocollimator based pointing control. A study was undertaken to assess the impact on the axis servos of the increased inertia and decreased frequencies and to provide a more complete model for the new servo design. This article describes the methodology of combining the dynamics of the structure with those of the servo and cites results for both the 64-m and 70-m antennas.

The development of a condensed antenna structural model [1] provides a control axis referenced representation of the structure dynamics in a compact form. When integrated with the dynamic properties of the hydraulic actuators and the control electronics, the result is a more comprehensive model for design, analysis, and simulation of the axis servos. The composite model is derived by coupling the linear differential equations describing the structure with those of the control components to produce a state variable model.

The structure model consists of a relatively large residual inertia to which are coupled individual modal inertias. For the elevation axis, two modes of the alidade structure are added. The elasticity of the control actuators, the compliance and inertia of the hydraulic system, the drive motors, and the characteristics of the servo compensation networks are also superimposed. In the most inclusive of the optional forms available, the elevation model representing five modal inertias and two alidade modes results in a nineteenth order linear system model.

The antenna pointing system employs two modes of position feedback derived from either a shaft angle encoder driven through a precision gear reducer from the bullgear, or an optical autocollimator mounted at the rear of the apex of the main reflector. The autocollimator mirror is attached to an hour

angle-declination mount on a pedestal isolated from that of the antenna. The structure model includes the transformation coefficients relating the displacement at the autocollimator reference structure to the arbitrary coordinates of the individual inertias. These coefficients enable system modeling of displacements at the autocollimator as well as at the encoder and tachometers, all of which are connected by nonzero flexibilities.

The state variable representation has one limitation in that system response properties are not evident by inspection when large system matrices are involved. For example, the existence of a pole in the right-half s plane, evident in a root locus display, is not recognizable by inspection of a linear system matrix unless the matrix happens to be in a form where the eigenvalues are recognizable. The matrix form of system representation does, however, provide great flexibility to accommodate various computer processing methods. Thus, both transfer function and block diagram representations are readily derived from the open loop as well as the closed rate loop system matrices. The results are available in a variety of forms suitable for classical as well as modern control system design and analysis.

An interactive Fortran program was developed for generating the model in state variable (matrix) form. Operating on an IBM PC, the program provides options for adjustment of the complexity of the model and of various parameter values. It displays the resultant linear system matrix and computes the corresponding system poles from the eigenvalues of the system matrix. Zeros of each of the tachometer, encoder, and autocollimator responses are determined from the eigenvalues of partitioned matrices formed according to the Mason gain formula [2]. Zeros of the autocollimator response are computed from the weighted sum of encoder and individual mode zeros where the weighting factors are the transformation coefficients mentioned earlier. This indirect method was used as an expedient to avoid coding an algorithm for the more general solution of zeros of the transformed matrix.

II. Structure Dynamics Model

The form of the condensed structure mode for the elevation axis is shown schematically in Fig. 1. In Fig. 1 θ_{A1}, θ_M, and θ_B correspond, respectively, to the angular rates of motion at the gear reducer attachment, the reducer output pinion (normalized with respect to the pinion to bullgear ratio), and the twin bullgears. The combined stiffness of the four gear reducers (two driving each bullgear) is represented by a single spring, K_G, shown connected between the pinion and the "single" bullgear. The angular rates of the individual modal inertias θ_1 through θ_N have an indirect correspondence to the motion of the physical antenna. A set of dimensionless coefficients, alpha, relate the motion of the Intermediate Reference Structure to θ_1 through θ_N.

While the azimuth axis of the antenna employs a single, stationary bullgear and moving gear reducers, its dynamic motion can be described using the same model as in the elevation axis. A small error results, however, because in azimuth, the reducer housing rotates with the alidade, such that in 360 degrees of azimuth motion each azimuth drive motor shaft makes $N-1$ revolutions relative to the pedestal and only N revolutions relative to the alidade. The error arises because the motor torque-speed characteristic is referenced to alidade coordinates, while acceleration torque is proportional to motion in inertial coordinates. However, since the gear ratio, N, is large, the resulting error is negligible.

Using the coordinate definitions of Fig. 1, the equations of motion for the individual inertias, J_i, are derived. In general, a small damping, D_i, is associated with each spring, K_i.

$$J_i\ddot{\theta}_i + D_i(\dot{\theta}_i - \dot{\theta}_B) + K_i(\theta_i - \theta_B) = 0 \qquad \text{for } i = 1 \text{ to } N \tag{1}$$

$$J_B\ddot{\theta}_B + (K_G + \sum_1^N K_i)\theta_B - \sum_1^N (D_i(\dot{\theta}_i - \dot{\theta}_B) + K_i\theta_i) = K_G\theta_M \tag{2}$$

$$J_{A1}\ddot{\theta}_{A1} + K_{A1}\theta_{A1} - K_{A1}\theta_{A2} + K_G(\theta_M - \theta_B) = 0 \tag{3}$$

$$J_{A2}\ddot{\theta}_{A2} + (K_{A1} + K_{A2})\theta_{A2} - K_{A1}\theta_{A1} = 0 \tag{4}$$

III. Hydraulic Motor Model

A schematic representation of the hydraulic system and gear reducers is shown in Fig. 2. Each gear reducer includes two fixed displacement, axial piston hydraulic motors driving a spur gear reducer. One motor provides control and the other produces a constant value countertorque. The countertorque motors are coupled to a regulated hydraulic pressure supply and connected so as to produce opposite torque outputs from two of the four reducers. This arrangement eliminates backlash in all gear meshes except the single mesh between each control motor and the first intermediate gear. The countertorque has the added benefit of increasing the incremental torque stiffness of the reducer geartrains. Control motor modulation is provided by a four-port hydraulic servovalve interconnecting a regulated high pressure hydraulic source and return with the motors.

The dynamic model of the drive motors and hydraulic system is derived according to established practices. Because the high gain of the rate servo loop diminishes the overall effects of friction, leakage, and valve pressure variations, a comparatively simple, second order, linearized model can be justified. Neglecting leakage, the oil flow, Q_m, through the hydraulic motor is given by

$$Q_m = V\dot{\theta}_m$$

where V is the volumetric displacement of the motor and $\dot{\theta}_m$ is the rotational speed at the output shaft.

Equating the fluid input power and the mechanical output power yields the relationship between motor torque, T_m, and hydraulic pressure, P

$$T_m = PV$$

The oil flow through the hydraulic servo valve is described by the equation of flow through a sharp edge orifice of variable size

$$Q_v = K_v I_v P_v^{1/2} \tag{5}$$

where P_v is the pressure drop across the valve, K_v is the flow constant of the valve, and I_v is the valve coil current.

The valve pressure drop equals the pressure difference between the regulated supply and the motor. Since both analytic and field test results show the steady state pressure drop is relatively constant in the range of 2000 to 2500 psi (13790 to 17238 kPa), a piecewise linear approximation of the above orifice equation is sufficient. Transient pressure drops caused by large acceleration or high wind torque will effectively decrease the rate loop gain and increase valve damping relative to the values represented by the linear approximation. With the aid of a root locus diagram, it can be shown that the net effect on the rate loop stability is negligible.

The linearized valve equation combined with the motor and compressible flow equations is illustrated in block diagram form in Fig. 3 where C and J_m represent the hydraulic compliance and motor inertia, respectively. The valve gain and damping constants K_p and D, respectively, are derived by partial differentiation of the valve flow Eq. (5) with values of P_v and Q_v that represent mean operating conditions. The damping parameter D may be increased to include motor leakage and other equivalent sources of damping.

In rearranging the above equations to a form compatible with the structure equations, a provision for the motion of the flexible structure supporting the reducer is required. This is accomplished by equating the motor rotation to the difference between θ_m and θ_{A1} in Fig. 1. Thus by application of the Mason gain rule to the diagram of Fig. 3 and with the definition of the reducer natural frequency

$$\omega_m = \left(\frac{V^2}{J_m C}\right)^{1/2}$$

the gear reducer equation is obtained as:

$$N_r\left(s^2 + \frac{D}{C}s + \omega_m^2\right)(\dot{\theta}_m - \dot{\theta}_A) + \frac{T_x}{N_r J_m}\left(s + \frac{D}{C}\right) = \frac{Q}{V}\omega_m^2 \tag{6}$$

where $Q = K_p I_v$ is the effective no-load valve flow and N_r is the overall motor-to-axis gear ratio.

The large value of gear ratio N_r (28700) permits the omission of the acceleration torque term that arises from the difference between reducer and inertial coordinates, with a small error resulting.

IV. System Equations

Equating the reducer shaft torque, T_x, to the torque transmitted by the reducer stiffness, K_G completes the equations defining the coupled actuation-structure system as follows:

$$T_x = K_G(\theta_m - \theta_B) \tag{7}$$

To accommodate analyses of the servo rate loops as well as the position loops, and also to facilitate computation of transfer function parameters, the system equations are developed in two steps. First, the open loop linear system matrix and associated input and output vectors are derived from Eqs. (1) through (7) above. Subsequently, the closed loop system matrix is formed by augmentation of the open loop matrix by inclusion of the rate feedback and compensation gains. The system state equations are derived from Eqs. (1) through (7) above using the state variable definitions in Table 1. The variable definitions are generalized to accommodate either the azimuth or elevation axis and a variable number of structure modes, N. Because the azimuth structure model is based upon the assumption of a rigid pedestal, state variables corresponding to $\dot{\theta}_{A1}$ and $\dot{\theta}_{A2}$ are absent from the azimuth model.

It will be seen that with the exception of the third, $(2N+5)$th, and $(2N+9)$th variables, all rates and accelerations are relative to stationary coordinates. The third variable corresponds to the effective rotor-stator rotation of the hydraulic motor and of the tachometer. The $(2N+5)$ and $(2N+9)$ variables corre-

spond to the effective extension of the gear reducer torsional stiffness for the azimuth/elevation axes, respectively. This torsional extension is related to the reducer output torque through the stiffness parameter K_G.

The resulting state equations for the elevation axis are listed in generalized form in Table 2. Using

$$\dot{\mathbf{X}} = [F]X + [G]U \tag{8a}$$

$$\mathbf{Y} = [H]X \tag{8b}$$

with $\mathbf{X}$ the state vector, $\dot{X}$ the time derivative of $\mathbf{X}$, U the system input, and Y the output, the corresponding linear system matrices are listed in Table 3. The system block diagram is shown in Fig. 4. The azimuth equations, matrices, and block diagram are similar to their elevation counterparts, except the four alidade states are omitted.

V. Derivation of Transfer Functions

The linear transfer functions relating the inputs and outputs are expressed as ratios of factored polynomials in the Laplace operator, s. The respective numerator and denominator factors are of the form (s – zero) and (s – pole) and are related to the system matrices through the traditional equations expressing the conditions for the poles and zeros of the response.

For system poles (denominator factors):

$$[sI - F] = [0] \tag{9}$$

For system zeros (numerator factors):

$$\begin{bmatrix} (sI - F) & -G \\ H & 0 \end{bmatrix} \begin{bmatrix} X(s) \\ U(s) \end{bmatrix} = [0] \tag{10a}$$

Equation (9) is an eigenvalue equation and the system poles are the eigenvalues of the linear system matrix. Software for eigenvalue evaluation is available for DOS microcomputers. For the more complicated case of Eq. (10) for evaluating the zeros, a more convenient method based on the Mason gain rule was employed. An advantage of this alternate method is that it reduces the dimensions of the matrices to be processed, thus improving numerical accuracy.

The method is based upon an adaptation of Mason's [2] signal flow graph gain (transfer function) determination, to state variable representations. Mason relates transfer function denominator and numerator to "determinants" of flow graphs and of certain subgraphs. It can be shown that these determinants of flow graphs are identical to the determinants of corresponding matrices representing the equations of the graphs. Using this equivalence, the transfer function numerator thus becomes the determinant of the partitioned matrix formed by deleting from the system equations those state variables included along the forward path in a flow graph representation of the equations. Since forward paths are recognizable in system equations, this partitioning can be accomplished without actually constructing the graph. The numerator factors are thus determined from the eigenvalues of the partitioned matrix formed above.

The zeros for the autocollimator and elevation encoder output responses are computed by superposition of individual components of the respective responses. This superposition avoids both a coordinate transformation of the system matrix and also the relatively complicated application of the Mason rule to a transformed matrix. Because it involves weighted summations of characteristic polynomials with subsequent factoring, this method is subject to numerical inaccuracy as matrix size increases. Good accuracy has resulted for models including three structure modes. The accuracy of this method could be improved and its usefulness extended to more modes by frequency scaling in such a way as to reduce the numeric range of the polynomials. Using superposition, the general case for the encoder response characteristic polynomial, P_E, is given by

$$P_E = \frac{K_G}{J_B} P_B - \frac{K_G}{J_{A1}} P_A \tag{10b}$$

where the response zeros are the roots of P_E; P_B and P_A are the characteristic polynomials of the bullgear and alidade responses, respectively. They are the characteristic polynomials of submatrices formed according to the method described above. Equation (10b) can be applied to the azimuth axis by equating P_A to zero.

The complexity of the elevation autocollimator response is reduced by partitioning into two smaller systems with the results combined so as to avoid processing polynomials of high order. The validity of this simplification is evident from the bullgear response submatrix formed by deleting the state variables included in the forward path between input and the bullgear rate. This submatrix is formed from the **F** matrix of Table 3 by deletion of rows and columns 1 through 4. Since this submatrix consists of two diagonal blocks, with one block representing the alidade and the other representing the tipping structure, the eigenvalues are simply the combination of those of the individual blocks. The complete solution of the autocollimator response is the weighted superposition of the responses of the bullgear and of the individual inertias, all of which are derived from this block diagonal matrix. The zeros of the

autocollimator response are thus the roots of characteristic polynomial P_{AC} defined by Eqs. (10c) and (10d):

$$P_{AC} = (P_A)\left(a_0 P_{BB} + \sum_1^N a_i \omega_i^2 P_i\right) \tag{10c}$$

$$\mathbf{F_B} = \begin{bmatrix} \mathbf{F}_{BB} & 0 \\ 0 & \mathbf{F}_A \end{bmatrix} \tag{10d}$$

where polynomials P are derived from their corresponding **F** matrices having the same subscripts, $\mathbf{F}_B$ is the bullgear response submatrix described above, and the $\mathbf{F}_i$ matrices are derived from $\mathbf{F}_{BB}$ by row, column deletion by the aforementioned principle. Coefficients $a_0 \cdots a_N$ in Eq. (10c) are the transformation coefficients provided in the structure model.

The physical significance of this partitioning is explained by considering the individual uncoupled responses of the alidade and the tipping structure to a torque input applied at the elevation gear reducer. The frequencies of infinite compliance (i.e., zero stiffness) of the alidade are frequencies of bullgear response zeros and are not influenced by the tipping structure. Additional frequencies of zero bullgear response are the natural frequencies of the individual tipping structure modes and are not influenced by the presence of the alidade. The partitioning is thus valid and no loss of generality results from its use.

VI. Closed Rate Loop Model

Modeling of the closed rate loop configuration is accomplished by extending the open loop equations and matrices to include the rate feedback and loop compensation. In both the 64-m and 70-m rate loops, tachometer based feedback, a lag/lead network, and a lead/lag network are employed to obtain a high degree of stiffness at low frequency and a comparatively narrow noise bandwidth. Since the 50 Hz servovalve bandwidth is sufficiently wide to have negligible effect on loop dynamics, the rate feedback can be modeled by the two phase compensation networks and a gain parameter.

The transfer function of the tachometer-network-amplifier, servovalve cascade is:

$$\frac{Q(s)}{\dot{\theta}_M(s)} = \frac{K_r V}{\omega_m^2} \frac{(s+Z_1)}{(s+P_1)} \frac{(s+Z_z)}{(s+P_2)} \tag{11}$$

The two real poles of Eq. (11) result in two additional states with the following additional state equations:

$$\dot{x}_M = -K_r(Z_1 - P_1)x_3 - P_1 x_M \tag{12}$$

$$\dot{x}_{M+1} = -K_r(Z_2 - P_2)x_3 + (Z_2 - P_2)x_M - P_2 x_{M+1} \tag{13}$$

where

$M = 2N+6$ for azimuth and
$M = 2N+10$ for elevation

For the closed loop case the equation for $\dot{x}_4$ becomes

$$\dot{x}_4 = -(\omega_m^2 + K_r)x_3 - \frac{D}{C} x_4 + x_m + x_{m+1} \tag{14}$$

The closed rate loop linear system matrix is obtained by augmenting the open loop matrix of Table 3 according to Eqs. (12), (13), and (14) and is shown in Table 4.

The various transfer functions for the open and closed loop systems are compiled according to the following properties of linear systems:

(1) all transfer functions in a given system have identical poles.

(2) the zeros in any given transfer function are invariant with respect to gain changes in other parts of the system not touching the forward path under consideration.

Property (1) implies that the tachometer, encoder, and autocollimator transfer functions all have identical poles, thus eliminating a need for repeated computations. Because the loop closing gain is in a feedback path, property (2) allows the use of zeros computed for the open loop case in the closed loop transfer functions. The real zero at -2.2, appearing in the closed loop functions, results from an electronic compensation network in the forward path.

VII. Numerical Results

A generalized Fortran program for generating the linear system matrix and computing the corresponding poles and zeros was developed and executed on an IBM PC. The program provides options for selection of axis, display and adjustment of parameter values, and adjustment of the number of structure modes included. As the number of modes is adjusted downward, the inertia of the rejected modes is added to the residual inertia, thus maintaining accuracy of the total inertia. Results are written to disk-files in a format compatible with postprocessing programs.

The Fortran program was used to derive models for the azimuth and elevation axes of both 64-m and 70-m configurations. Hardware parameter values available for the four configurations are listed in Table 5. In some cases the number of structure modes modeled was reduced from the number available in order to reduce computation and data space. For the 64-m azimuth case the 4th mode was eliminated, for 70-m azimuth the 5th mode was eliminated, and for 70-m elevation the 3rd, 4th, and 5th modes were merged into a weighted composite. The best estimate of structure damping ratio available is 0.003, which was observed in factory tests of the 70-m quadripod. Because of computation error in some of the zeros computations resulting from nonzero damping, all computations were run with zero structure damping. The introduction of appropriate damping would displace the complex zeros from the imaginary axis and cause a similar displacement of the complex poles. However, since the poles are already damped by the rate loop, the introduction of structure damping would have small effect on the overall results.

The poles and zeros of the open and closed rate loop transfer functions are listed in Table 6. In cases where inertia values other than those of Table 5 were used, the actual values are included in Table 6. Frequency response plots of amplitude vs. radian frequency for the 70-m azimuth and elevation axes are shown in Figs. 5 and 6.

The differences between the tachometer, encoder, and autocollimator responses are due to the flexibility of the gear reducers and structure between the respective devices. In the elevation autocollimator responses, the low frequency (2 Hz for 64-m, 1.5 Hz for 70-m) resonant peak and subsequent roll-off result from the flexibility of the alidade. It occurs because alidade deflections in elevation are sensed by the autocollimator but not by the encoder. Both the encoder and autocollimator responses for both 64-m axes show a high frequency roll-off beginning at 6.7 to 7.2 Hz. For the 70-m configuration the roll-off frequencies are nearly identical, 6.5 to 6.8 Hz. This result is explained by the fact that most of the inertia increase resulting from the upgrade is associated with the first three modes while the residual inertias are relatively unchanged.

VIII. Conclusions

The models described have enabled the design of the rate and position servos for the 70-m configuration with a minimum of on-site adjustment. The azimuth axis structure model, see Table 5, includes three low frequency modes at 1.27 to 1.42 Hz, a frequency roughly half that of the lowest mode in the 64-m configuration. These modes are a cause for concern because of their low frequencies, close spacing, and relatively large associated inertias which make damping by the servo loop difficult to achieve. The presence of these modes necessitated an increase of the motor shaft mounted inertia wheels as a means of diminishing their effect on the control system.

The results summarized in Table 6 provide improved definition of the encoder and autocollimator response characteristics as compared with those based on the assumption of a "rigid" alidade structure. These results will be useful in future efforts to improve autocollimator based pointing. They do not necessitate immediate changes as a result of the upgrade.

References

[1] R. Levy, "Condensed Antenna Structural Models for Dynamic Analysis," *TDA Progress Report 42-80*, vol. October–December 1984, Jet Propulsion Laboratory, Pasadena, California, pp. 40–61, February 15, 1985.

[2] S. J. Mason, "Feedback Theory-Further Properties of Signal Flow Graphs," *Proc. IRE*, vol. 44, pp. 920–926, July 1956.

Table 1. State variable definitions

Variable	Symbol	Description
x_1	$\dot{\theta}_B$	Bullgear angular rate
x_2	$\dot{x}_1$	Bullgear acceleration
x_3	$\dot{\theta}_M - \dot{\theta}_{A1}$	Motor/tach rate
x_4		Hydraulic torque/J_m
x_{2i+3}	$\dot{\theta}_i$	Angular rate at the inertia J_i for $i = 1$ to N
x_{2i+4}	$\dot{x}_{2i+3}$	
x_{2N+5}	$\theta_M - \theta_B$	Motor–bullgear angle difference, azimuth only
x_{2N+5}	$\dot{\theta}_{A1}$	Angular rate at alidade 1, elevation only
x_{2N+6}	$\dot{x}_{2N+5}$	Elevation only
x_{2N+7}	$\dot{\theta}_{A2}$	Angular rate at alidade 2, elevation only
x_{2N+8}	$\dot{x}_{2N+7}$	Elevation only
x_{2N+9}	$\theta_M - \theta_B$	Motor–bullgear angle difference, elevation only

Table 2. State equations for elevation

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -\left(\frac{K_G}{J_B} + \sum_1^N \frac{K_i}{J_B}\right) x_1 + \sum_1^N \frac{K_i}{J_i} x_{2i+3} + \frac{K_G}{J_B} x_3 + \frac{K_G}{J_B} x_{2N+5}$$

$$\dot{x}_3 = x_4 - \frac{K_G}{J_M} x_{2N+9}$$

$$\dot{x}_4 = -\omega_m^2 x_3 - \frac{D}{C} x_4 + \frac{Q}{V} \omega_m^2$$

$$\dot{x}_{2i+3} = x_{2i+4} \qquad \text{for } i = 1 \text{ to } N$$

$$\dot{x}_{2i+4} = \omega_i^2 x_i \neq \frac{D_i}{J_i} x_2 - \omega_i^2 x_{2i+3} - \frac{D_i}{J_i} x_{2i+4}$$

$$\vdots$$

$$\dot{x}_{2N+5} = x_{2N+6}$$

$$\dot{x}_{2N+6} = \frac{K_G}{J_{A1}} x_1 - \frac{K_G}{J_{A1}} x_3 - \frac{K_G + K_{A1}}{J_{A1}} x_{2N+5} + \frac{K_{A1}}{J_{A1}} x_{2N+7}$$

$$\dot{x}_{2N+7} = x_{2N+8}$$

$$\dot{x}_{2N+8} = \frac{K_{A1}}{J_{A2}} x_{2N+5} - \frac{K_{A1} + K_{A2}}{J_{A2}} x_{2N+7}$$

$$\dot{x}_{2N+9} = -x_1 + x_3 + x_{2N+5}$$

N = number of structure modes in model

Table 3. Linear system matrices for open loop elevation axis

$$
F = \begin{bmatrix}
0 & 1 & & & & & & & & & & & & & 0 \\
\frac{-K_G - K_T}{J_B} & 0 & \frac{K_G}{J_B} & 0 & \frac{K_1}{J_1} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{K_G}{J_B} & & & 0 \\
 & & 0 & 1 & & & & & & & & & & & \frac{-K_G}{J_M} \\
 & & -\omega_m^2 & \frac{-D}{C} & & & & & & & & & & & \\
 & & & 1 & 0 & & & & & & & & & & \\
\omega_1^2 & \frac{D_1}{J_1} & & & -\omega_1^2 & \frac{-D_1}{J_1} & & & & & & & & & \\
 & & & & & & 0 & & & & & & & & \\
 & & & & & & & 1 & 0 & & & & & & \\
\omega_N^2 & \frac{D_N}{J_N} & & & & & & & -\omega_N^2 & \frac{-D_N}{J_N} & & & & & \\
 & & & & & & & & & & & 0 & 1 & & \\
\frac{K_G}{J_{A1}} & 0 & \frac{-K_G}{J_{A1}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{K_G + K_{A1}}{J_{A1}} & 0 & \frac{K_{A1}}{J_{A1}} & \\
 & & & & & & & & & & & & & 0 & 1 \\
 & & & & & & & & & & & \frac{K_{A1}}{J_{A2}} & 0 & -\frac{K_{A1} + K_{A2}}{J_{A2}} & 0 \\
-1 & 0 & 1 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 1 & 0 & & 0
\end{bmatrix}
$$

Note: $K_T = \sum_1^N K_i$

LINEAR SYSTEM INPUT MATRIX, G

$$
G = \left(\frac{\omega_m^2}{V}\right) \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}
$$

Table 3. (contd)

OUTPUT VECTORS, H

FOR TACHOMETER OUTPUT RESPONSE

$\mathbf{H_T}$ = [0 0 1 0 ..... 0]

FOR ENCODER OUTPUT RESPONSE

$\mathbf{H_{Ea}}$ = [1 0 0 0 ..... 0] Azimuth only

$\mathbf{H_{Ee}}$ = [1 0 0 0 ... −1 .... 0] Elevation only

FOR AUTOCOLLIMATOR MOUNT RESPONSE

$\mathbf{H_{AC}}$ = [a0 0 0 0 a1 0 a2 0 a3 0 a4 0 a5 0 .... 0]

NOTES:

All input and output vectors have lengths equal to the dimension of the corresponding linear system matrix, F.

The −1 in the Elevation Encoder response is in column $2N + 5$, where N is the number of structure modes modeled.

Coefficients a1. . .a5 are defined as a part of the structure model. When fewer than five structure modes are modeled the a's corresponding to unmodeled modes are replaced with zero and a0 is increased an equal amount such that the sum of a0 + a1 + . . . aN = 1 where N is the number of structure modes modeled.

Table 4. Linear system matrices for closed rate loop

$$F_{CL} = \begin{bmatrix} \begin{bmatrix} & & \\ & F - K_r T & \\ & & \end{bmatrix} & \begin{matrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \end{matrix} & \begin{matrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \end{matrix} \\ \begin{matrix} 0 & 0 & -K_r(Z_2 - P_2) & 0 & \cdots \end{matrix} & -P_2 & (Z_2 - P_2) \\ \begin{matrix} 0 & 0 & -K_r(Z_1 - P_1) & 0 & \cdots \end{matrix} & 0 & -P_1 \end{bmatrix}$$

where **T** is a matrix with 1.0 in Row 4, Col. 3 and zeros elsewhere

$$G_{CL} = \frac{\omega_m^2}{\nu} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ Z_2 - P_2 \\ 0 \end{bmatrix}$$

H_T, H_E, H_{AC} are formed from their open loop counterparts by adding two zero elements following the last element

Table 5. 64- and 70-m antenna structure and servo parameter values

	64-m AZ	70-m AZ	64-m EL	70-m EL
RESIDUAL (Base) INERTIA, J_B	0.1783 (0.2418)	0.1813 (0.2459)	0.0840 (0.1139)	0.1066 (0.1446)
INERTIA Ratio, J_1/J_B	0.3097	0.2356	0.3782	1.1746
INERTIA Ratio, J_2/J_B	0.1092	0.4368	0.3936	0.2383
INERTIA Ratio, J_3/J_B	0.0505	0.2494	0.2729	0.0433
INERTIA Ratio, J_4/J_B	0.0975	0.0436		0.0812
INERTIA Ratio, J_5/J_B		0.0132		0.0121
MOTOR INERTIA, J_M	0.6640 (0.9006)	1.0000 (1.3563)	0.6640 (0.9006)	0.6640 (0.9006)
ALIDADE INERTIA, J_{A1}			0.0145 (0.0197)	0.0145 (0.0197)
ALIDADE INERTIA, J_{A2}			0.1330 (0.1804)	0.1330 (0.1804)
FREQ of $(Kgear/J_B)^{0.5}$	38.3639	38.0399	66.5865	59.1247
FREQ, mode 1, R/s	15.0500	7.9670	19.3830	14.7910
FREQ, mode 2, R/s	25.7600	8.3120	20.7600	17.6930
FREQ, mode 3, R/s	32.0800	9.8900	26.4020	20.2760
FREQ, mode 4, R/s	36.0200	13.7470		21.8340
FREQ, mode 5, R/s		16.9020		25.6290
FREQ of Hyd Motor	9.5940	7.8000	9.5940	9.5940
RE part, Motor freq, D/C	1.2000	1.2000	1.2000	1.2000
FREQ of Alidade 1			67.8258	67.8258
FREQ of Alidade 2			32.0237	32.0237
IRS Transformation coeff. a1	0.1318	0.1331	0.1421	0.2939
a2	0.2022	0.2767	0.1722	0.0977
a3	0.1040	0.1169	0.2090	0.0266
a4	0.0618	0.0099		0.0309
a5		0.0376		0.0050
RATE LOOP GAIN, K_R/ω_m^2	2302	2302	2302	2302

RATE LOOP COMPENSATION NETWORKS, all axes

$$\frac{(s + 2.2)}{(s + 0.12)} \quad , \quad \frac{(s + 7.1)}{(s + 81)}$$

NOTES:

64-m and 70-m azimuth models are for 90 degree elevation.

Inertia values are ft-lb-s^2, (kg-M^2) and are referred to motor shaft.

Natural frequencies are in radians/s.

Coefficient a0 = 1.0000 – a1 – a2 – a3 – a4 – a5.

Table 6a. Transfer functions for 64-m azimuth axis: 3 structure modes modeled

(FILE: AZ6415)	
RESIDUAL (Base) INERTIA, JB	0.1956 (0.2653)
INERTIA Ratio, J1/JB	0.2821
INERTIA Ratio, J2/JB	0.0995
INERTIA Ratio, J3/JB	0.0460
INERTIA Ratio, J4/JB	0.0000
INERTIA Ratio, J5/JB	0.0000
MOTOR INERTIA, JM	0.6640 (0.9006)

OPEN RATE LOOP

VALVE CURRENT TO MOTOR SHAFT TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-5.0933E-003	4.4378E+001	0.0000E+000	4.0634E+001
-5.0933E-003	-4.4378E+001	0.0000E+000	-4.0634E+001
-1.7687E-003	3.1629E+001	0.0000E+000	3.1012E+001
-1.7687E-003	-3.1629E+001	0.0000E+000	-3.1012E+001
-4.2904E-003	2.5545E+001	0.0000E+000	2.4712E+001
-4.2904E-003	-2.5545E+001	0.0000E+000	-2.4712E+001
-2.3554E-002	1.5505E+001	0.0000E+000	1.4625E+001
-2.3554E-002	-1.5505E+001	0.0000E+000	-1.4625E+001
-5.6529E-001	7.8390E+000		
-5.6529E-001	-7.8390E+000		

CLOSED RATE LOOP

RATE LOOP INPUT TO ENCODER RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.1927E+000	4.4833E+001		
-3.1927E+000	-4.4833E+001		
-5.5539E-001	3.1564E+001	0.0000E+000	3.2080E+001
-5.5539E-001	-3.1564E+001	0.0000E+000	-3.2080E+001
-8.6159E-001	2.5261E+001	0.0000E+000	2.5760E+001
-8.6159E-001	-2.5261E+001	0.0000E+000	-2.5760E+001
-5.3122E-001	1.4465E+001	0.0000E+000	1.5050E+001
-5.3122E-001	-1.4465E+001	0.0000E+000	-1.5050E+001
-1.2952E+001	1.6448E+001		
-1.2952E+001	-1.6448E+001		
-1.4483E+000	0.0000E+000	-2.2000E+000	0.0000E+000
-4.4686E+001	0.0000E+000	-8.1000E+001	0.0000E+000

CLOSED RATE LOOP

RATE LOOP INPUT TO TACHOMETER RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.1927E+000	4.4833E+001	0.0000E+000	4.0634E+001
-3.1927E+000	-4.4833E+001	0.0000E+000	-4.0634E+001
-5.5539E-001	3.1564E+001	0.0000E+000	3.1012E+001
-5.5539E-001	-3.1564E+001	0.0000E+000	-3.1012E+001
-8.6159E-001	2.5261E+001	0.0000E+000	2.4712E+001
-8.6159E-001	-2.5261E+001	0.0000E+000	-2.4712E+001
-5.3122E-001	1.4465E+001	0.0000E+000	1.4625E+001
-5.3122E-001	-1.4465E+001	0.0000E+000	-1.4625E+001
-1.2952E+001	1.6448E+001		
-1.2952E+001	-1.6448E+001		
-1.4483E+000	0.0000E+000	-2.2000E+000	0.0000E+000
-4.4686E+001	0.0000E+000	-8.1000E+001	0.0000E+000

CLOSED RATE LOOP

RATE LOOP INPUT TO AUTOCOLLIMATOR RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.1927E+000	4.4833E+001		
-3.1927E+000	-4.4833E+001		
-5.5539E-001	3.1564E+001	0.0000E+000	3.6678E+001
-5.5539E-001	-3.1564E+001	0.0000E+000	-3.6678E+001
-8.6159E-001	2.5261E+001	0.0000E+000	2.8297E+001
-8.6159E-001	-2.5261E+001	0.0000E+000	-2.8297E+001
-5.3122E-001	1.4465E+001	0.0000E+000	1.5985E+001
-5.3122E-001	-1.4465E+001	0.0000E+000	-1.5985E+001
-1.2952E+001	1.6448E+001		
-1.2952E+001	-1.6448E+001		
-1.4483E+000	0.0000E+000	-2.2000E+000	0.0000E+000
-4.4686E+001	0.0000E+000	-8.1000E+001	0.0000E+000

Table 6b. Transfer functions for 70-m azimuth axis: 4 structure modes modeled

(FILE: AZ7028)	
RESIDUAL (Base) INERTIA, JB	0.1837 (0.2492)
INERTIA Ratio, J1/JB	0.2325
INERTIA Ratio, J2/JB	0.4311
INERTIA Ratio, J3/JB	0.2462
INERTIA Ratio, J4/JB	0.0430
INERTIA Ratio, J5/JB	0.0000
MOTOR INERTIA, JM	1.0000 (0.1356)

OPEN RATE LOOP

VALVE CURRENT TO MOTOR SHAFT TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.1307E-003	4.2053E+001	0.0000E+000	3.8858E+001
-3.1307E-003	-4.2053E+001	0.0000E+000	-3.8858E+001
-2.8774E-003	1.3785E+001	0.0000E+000	1.3706E+001
-2.8774E-003	-1.3785E+001	0.0000E+000	-1.3706E+001
-6.1942E-002	1.0306E+001	0.0000E+000	9.8101E+000
-6.1942E-002	-1.0306E+001	0.0000E+000	-9.8101E+000
-7.6929E-002	8.9059E+000	0.0000E+000	8.2325E+000
-7.6929E-002	-8.9059E+000	0.0000E+000	-8.2325E+000
-5.9236E-003	8.0631E+000	0.0000E+000	7.9107E+000
-5.9236E-003	-8.0631E+000	0.0000E+000	-7.9107E+000
-4.4911E-001	6.1701E+000		
-4.4911E-001	-6.1701E+000		

CLOSED RATE LOOP

RATE LOOP INPUT TO ENCODER RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-1.5156E+000	4.2566E+001		
-1.5156E+000	-4.2566E+001		
-4.8021E-002	1.3696E+001	0.0000E+000	1.3747E+001
-4.8021E-002	-1.3696E+001	0.0000E+000	-1.3747E+001
-6.7226E-002	9.7472E+000	0.0000E+000	9.8900E+000
-6.7226E-002	-9.7472E+000	0.0000E+000	-9.8900E+000
-3.6465E-002	8.1685E+000	0.0000E+000	8.3120E+000
-3.6465E-002	-8.1685E+000	0.0000E+000	-8.3120E+000
-9.1012E-002	7.8285E+000	0.0000E+000	7.9670E+000
-9.1012E-002	-7.8285E+000	0.0000E+000	-7.9670E+000
-8.4011E+000	1.3145E+001		
-8.4011E+000	-1.3145E+001		
-6.0554E+001	0.0000E+000	-8.1000E+001	0.0000E+000
-1.4472E+000	0.0000E+000	-2.2000E+000	0.0000E+000

CLOSED RATE LOOP

RATE LOOP INPUT TO TACHOMETER RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-1.5156E+000	4.2566E+001	0.0000E+000	3.8858E+001
-1.5156E+000	-4.2566E+001	0.0000E+000	-3.8858E+001
-4.8021E-002	1.3696E+001	0.0000E+000	1.3706E+001
-4.8021E-002	-1.3696E+001	0.0000E+000	-1.3706E+001
-6.7226E-002	9.7472E+000	0.0000E+000	9.8101E+000
-6.7226E-002	-9.7472E+000	0.0000E+000	-9.8101E+000
-3.6465E-002	8.1685E+000	0.0000E+000	8.2325E+000
-3.6465E-002	-8.1685E+000	0.0000E+000	-8.2325E+000
-9.1012E-002	7.8285E+000	0.0000E+000	7.9107E+000
-9.1012E-002	-7.8285E+000	0.0000E+000	-7.9107E+000
-8.4011E+000	1.3145E+001		
-8.4011E+000	-1.3145E+001		
-6.0554E+001	0.0000E+000	-8.1000E+001	0.0000E+000
-1.4472E+000	0.0000E+000	-2.2000E+000	0.0000E+000

CLOSED RATE LOOP

RATE LOOP INPUT TO AUTOCOLLIMATOR RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-1.5156E+000	4.2566E+001		
-1.5156E+000	-4.2566E+001		
-4.8021E-002	1.3696E+001	0.0000E+000	1.3353E+001
-4.8021E-002	-1.3696E+001	0.0000E+000	-1.3353E+001
6.7226E-002	9.7472E+000	0.0000E+000	1.3009E+001
-6.7226E-002	-9.7472E+000	0.0000E+000	-1.3009E+001
-3.6465E-002	8.1685E+000	0.0000E+000	9.2957E+000
-3.6465E-002	-8.1685E+000	0.0000E+000	-9.2957E+000
-9.1012E-002	7.8285E+000	0.0000E+000	8.0628E+000
-9.1012E-002	-7.8285E+000	0.0000E+000	-8.0628E+000
-8.4011E+000	1.3145E+001		
-8.4011E+000	-1.3145E+001		
-6.0554E+001	0.0000E+000	-8.1000E+001	0.0000E+000
-1.4472E+000	0.0000E+000	-2.2000E+000	0.0000E+000

Table 6c. Transfer functions for 64-m elevation axis: 3 tipping structure and 2 alidade modes modeled

(FILE: EL6414)	
RESIDUAL (Base) INERTIA, JB	0.0840 (0.1139)
INERTIA Ratio, J1/JB	0.3782
INERTIA Ratio, J2/JB	0.3936
INERTIA Ratio, J3/JB	0.2729
INERTIA Ratio, J4/JB	0.0000
INERTIA Ratio, J5/JB	0.0000
MOTOR INERTIA, JM	0.6640 (0.9006)

OPEN RATE LOOP

VALVE CURRENT TO MOTOR SHAFT TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-2.2606E-005	1.8661E+002	0.0000E+000	1.8528E+002
-2.2606E-005	-1.8661E+002	0.0000E+000	-1.8528E+002
-5.2814E-004	4.3688E+001	0.0000E+000	4.3309E+001
-5.2814E-004	-4.3688E+001	0.0000E+000	-4.3309E+001
-3.8028E-005	3.1688E+001	0.0000E+000	3.1678E+001
-3.8028E-005	-3.1688E+001	0.0000E+000	-3.1678E+001
-2.9061E-004	2.3835E+001	0.0000E+000	2.3806E+001
-2.9061E-004	-2.3835E+001	0.0000E+000	-2.3806E+001
-1.7055E-004	1.9984E+001	0.0000E+000	1.9976E+001
-1.7055E-004	-1.9984E+001	0.0000E+000	-1.9976E+001
-9.0286E-002	1.4827E+001	0.0000E+000	1.2709E+001
-9.0286E-002	-1.4827E+001	0.0000E+000	-1.2709E+001
-5.0871E-001	8.0624E+000		
-5.0871E-001	-8.0624E+000		

CLOSED RATE LOOP

RATE LOOP INPUT TO ENCODER RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.2473E-002	1.8668E+002		
-3.2473E-002	-1.8668E+002		
-2.9969E-001	4.3696E+001	0.0000E+000	4.4681E+001
-2.9969E-001	-4.3696E+001	0.0000E+000	-4.4681E+001
-9.6810E-003	3.1684E+001	0.0000E+000	3.1582E+001
-9.6810E-003	-3.1684E+001	0.0000E+000	-3.1582E+001
-2.2261E-002	2.3809E+001	0.0000E+000	2.3968E+001
-2.2261E-002	-2.3809E+001	0.0000E+000	-2.3968E+001
-4.3576E-003	1.9975E+001	0.0000E+000	1.9799E+001
-4.3576E-003	-1.9975E+001	0.0000E+000	-1.9799E+001
-4.9794E-001	1.2338E+001	0.0000E+000	1.3135E+001
-4.9794E-001	-1.2338E+001	0.0000E+000	-1.3135E+001
-2.0885E+001	1.6166E+001		
-2.0885E+001	-1.6166E+001		
-3.7370E+001	0.0000E+000	-8.1000E+001	0.0000E+000
-1.4486E+000	0.0000E+000	-2.2000E+000	0.0000E+000

CLOSED RATE LOOP

RATE LOOP INPUT TO TACHOMETER RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.2473E-002	1.8668E+002	0.0000E+000	1.8528E+002
-3.2473E-002	-1.8668E+002	0.0000E+000	-1.8528E+002
-2.9969E-001	4.3696E+001	0.0000E+000	4.3309E+001
-2.9969E-001	-4.3696E+001	0.0000E+000	-4.3309E+001
-9.6810E-003	3.1684E+001	0.0000E+000	3.1678E+001
-9.6810E-003	-3.1684E+001	0.0000E+000	-3.1678E+001
-2.2261E-002	2.3809E+001	0.0000E+000	2.3806E+001
-2.2261E-002	-2.3809E+001	0.0000E+000	-2.3806E+001
-4.3576E-003	1.9975E+001	0.0000E+000	1.9976E+001
-4.3576E-003	-1.9975E+001	0.0000E+000	-1.9976E+001
-4.9794E-001	1.2338E+001	0.0000E+000	1.2709E+001
-4.9794E-001	-1.2338E+001	0.0000E+000	-1.2709E+001
-2.0885E+001	1.6166E+001		
-2.0885E+001	-1.6166E+001		
-3.7370E+001	0.0000E+000	-8.1000E+001	0.0000E+000
-1.4486E+000	0.0000E+000	-2.2000E+000	0.0000E+000

CLOSED RATE LOOP

RATE LOOP INPUT TO AUTOCOLLIMATOR RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.2473E-002	1.8668E+002		
-3.2473E-002	-1.8668E+002		
-2.9969E-001	4.3696E+001	0.0000E+000	7.2267E+001
-2.9969E-001	-4.3696E+001	0.0000E+000	-7.2267E+001
-9.6810E-003	3.1684E+001	0.0000E+000	3.4188E+001
-9.6810E-003	-3.1684E+001	0.0000E+000	-3.4188E+001
-2.2261E-002	2.3809E+001	0.0000E+000	3.0055E+001
-2.2261E-002	-2.3809E+001	0.0000E+000	-3.0055E+001
-4.3576E-003	1.9975E+001	0.0000E+000	2.2681E+001
-4.3576E-003	-1.9975E+001	0.0000E+000	-2.2681E+001
-4.9794E-001	1.2338E+001	0.0000E+000	1.9844E+001
-4.9794E-001	-1.2338E+001	0.0000E+000	-1.9844E+001
-2.0885E+001	1.6166E+001		
-2.0885E+001	-1.6166E+001		
-3.7370E+001	0.0000E+000	-8.1000E+001	0.0000E+000
-1.4486E+000	0.0000E+000	-2.2000E+000	0.0000E+000

Table 6d. Transfer functions for 70-m elevation axis: 3 tipping structure and 2 alidade modes modeled

INERTIA AND FREQ OF 3rd MODE ARE ADJUSTED COMPOSITES OF MODES 3, 4, 5

(FILE: EL7012)	
RESIDUAL (Base) INERTIA, JB	0.1066 (0.1446)
INERTIA Ratio, J1/JB	1.1746
INERTIA Ratio, J2/JB	0.2383
INERTIA Ratio, J3/JB	0.1362
FREQ, mode 3, R/s	21.7400
INERTIA Ratio, J4/JB	0.0000
INERTIA Ratio, J5/JB	0.0000
MOTOR INERTIA, JM	0.6640 (0.9006)

OPEN RATE LOOP

VALVE CURRENT TO MOTOR SHAFT TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-2.3597E-005	1.8433E+002	0.0000E+000	1.8299E+002
-2.3597E-005	-1.8433E+002	0.0000E+000	-1.8299E+002
-6.7820E-004	4.1970E+001	0.0000E+000	4.1540E+001
-6.7820E-004	-4.1970E+001	0.0000E+000	-4.1540E+001
-1.3184E-004	2.7651E+001	0.0000E+000	2.7629E+001
-1.3184E-004	-2.7651E+001	0.0000E+000	-2.7629E+001
-7.1847E-004	2.0753E+001	0.0000E+000	2.0711E+001
-7.1847E-004	-2.0753E+001	0.0000E+000	-2.0711E+001
-1.8263E-003	1.7085E+001	0.0000E+000	1.7039E+001
-1.8263E-003	-1.7085E+001	0.0000E+000	-1.7039E+001
-1.8590E-001	1.2952E+001	0.0000E+000	9.8583E+000
-1.8509E-001	-1.2952E+001	0.0000E+000	-9.8583E+000
-4.1066E-001	7.1233E+000		
-4.1066E-001	-7.1233E+000		

CLOSED RATE LOOP

RATE LOOP INPUT TO ENCODER RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.3846E-002	1.8441E+002		
-3.3846E-002	-1.8441E+002		
-3.5693E-001	4.1960E+001	0.0000E+000	4.2889E+001
-3.5693E-001	-4.1960E+001	0.0000E+000	-4.2889E+001
-2.0608E-002	2.7636E+001	0.0000E+000	2.7644E+001
-2.0608E-002	-2.7636E+001	0.0000E+000	-2.7644E+001
-2.5872E-002	2.0708E+001	0.0000E+000	2.0737E+001
-2.5872E-002	-20.708E+001	0.0000E+000	-2.0737E+001
-1.5653E-002	1.7033E+001	0.0000E+000	1.7050E+001
-1.5653E-002	-1.7033E+001	0.0000E+000	-1.7050E+001
-3.2123E-001	9.4835E+000	0.0000E+000	1.0185E+001
-3.2123E-001	-9.4835E+000	0.0000E+000	1.0185E+001
-2.0812E+001	1.6437E+001		
-2.0812E+001	-1.6437E+001		
-3.7699E+001	0.0000E+000	-8.1000E+001	0.0000E+000
-1.4483E+000	0.0000E+000	-2.2000E+000	0.0000E+000

CLOSED RATE LOOP

RATE LOOP INPUT TO TACHOMETER RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.3846E-002	1.8441E+002	0.0000E+000	1.8299E+002
-3.3846E-002	-1.8441E+002	0.0000E+000	-1.8299E+002
-3.5693E-001	4.1960E+001	0.0000E+000	4.1540E+001
-3.5693E-001	-4.1960E+001	0.0000E+000	-4.1540E+001
-2.0608E-002	2.7636E+001	0.0000E+000	2.7629E+001
-2.0608E-002	-2.7636E+001	0.0000E+000	-2.7629E+001
-2.5872E-002	2.0708E+001	0.0000E+000	2.0711E+001
-2.5872E-002	-2.0708E+001	0.0000E+000	-2.0711E+001
-1.5653E-002	1.7033E+001	0.0000E+000	1.7039E+001
-1.5653E-002	-1.7033E+001	0.0000E+000	-1.7039E+001
-3.2123E-001	9.4835E+000	0.0000E+000	9.8583E+000
-3.2123E-001	-9.4835E+000	0.0000E+000	-9.8583E+000
-2.0812E+001	1.6437E+001		
-2.0812E+001	-1.6437E+001		
-3.7699E+001	0.0000E+000	-8.1000E+001	0.0000E+000
-1.4483E+000	0.0000E+000	-2.2000E+000	0.0000E+000

CLOSED RATE LOOP

RATE LOOP INPUT TO AUTOCOLLIMATOR RATE TRANSFER FUNCTION

POLES		ZEROS	
Real	Imag	Real	Imag
-3.3846E-002	1.8441E+002		
-3.3846E-002	-1.8441E+002		
-3.5693E-001	4.1960E+001	0.0000E+000	7.2267E+001
-3.5693E-001	-4.1960E+001	0.0000E+000	-7.2267E+001
-2.0608E-002	2.7636E+001	0.0000E+000	3.0055E+001
-2.0608E-002	-2.7636E+001	0.0000E+000	-3.0055E+001
-2.5872E-002	2.0708E+001	0.0000E+000	2.2960E+001
-2.5872E-002	-2.0708E+001	0.0000E+000	-2.2960E+001
-1.5653E-002	1.7033E+001	0.0000E+000	1.9858E+001
-1.5653E-002	-1.7033E+001	0.0000E+000	-1.9858E+001
-3.2123E-001	9.4835E+000	0.0000E+000	1.6359E+001
-3.2123E-001	-9.4835E+000	0.0000E+000	-1.6359E+001
-2.0812E+001	1.6437E+001		
-2.0812E+001	-1.6437E+001		
-3.7699E+001	0.0000E+000	-8.1000E+001	0.0000E+000
-1.4483E+000	0.0000E+000	-2.2000E+000	0.0000E+000

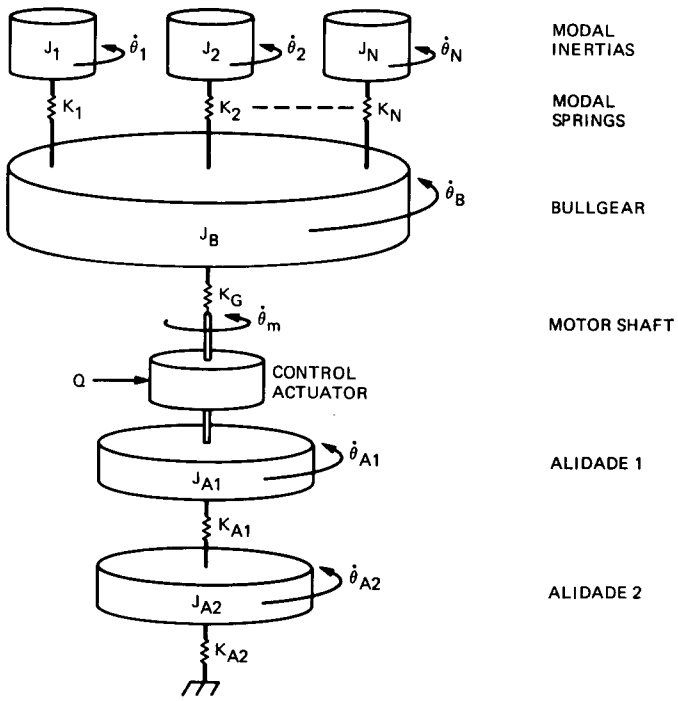


Fig. 1. Structure model schematic representation

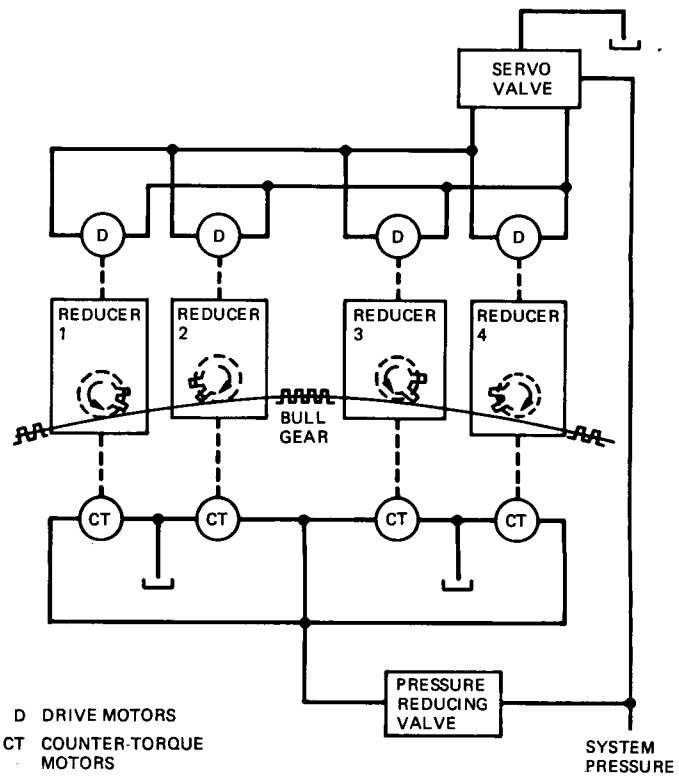


Fig. 2. Countertorque system, 64/70-m antenna

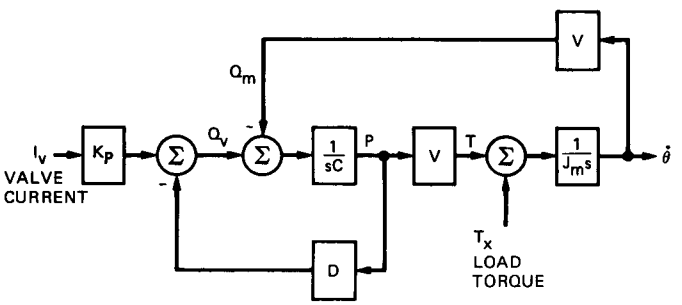


Fig. 3. Hydraulic motor linearized block diagram

C-4

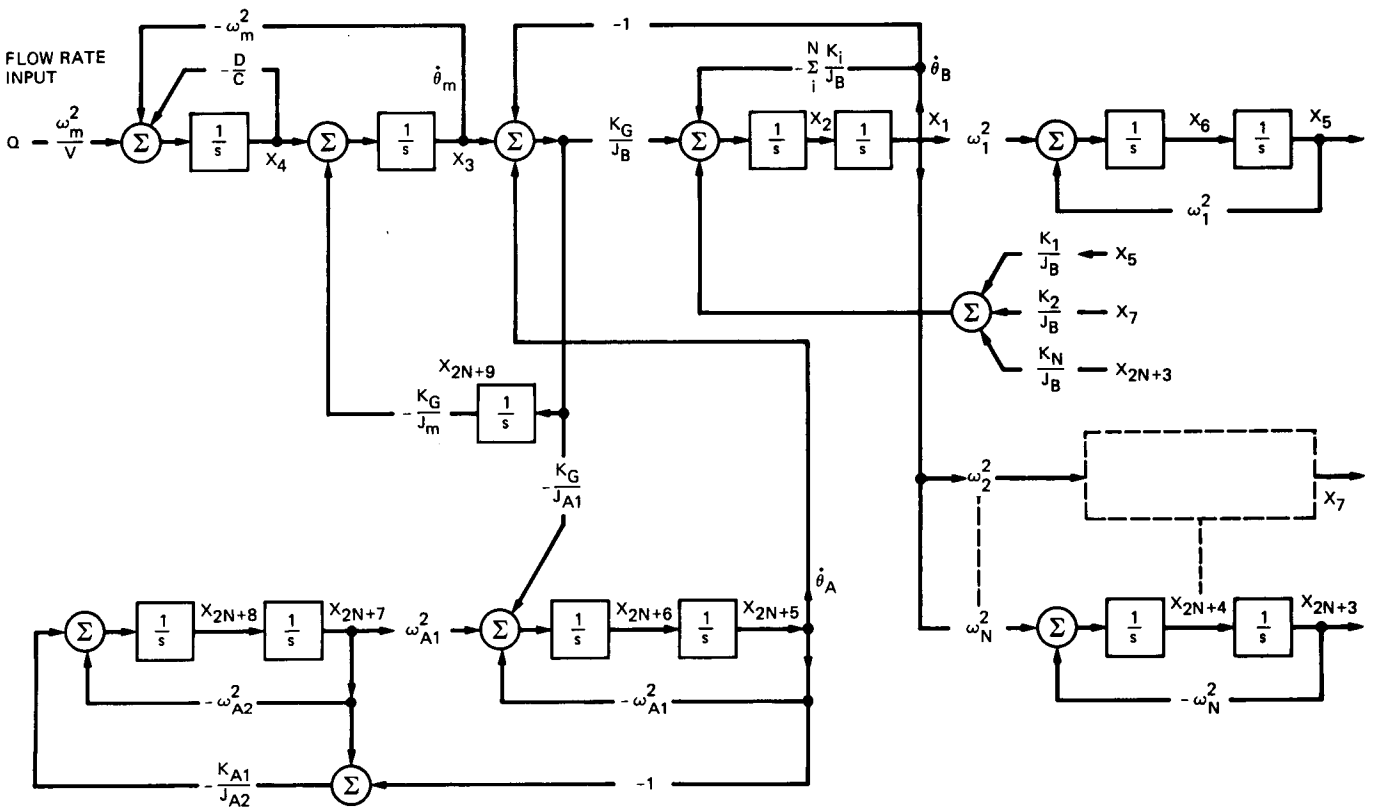


Fig. 4. Elevation open loop system block diagram

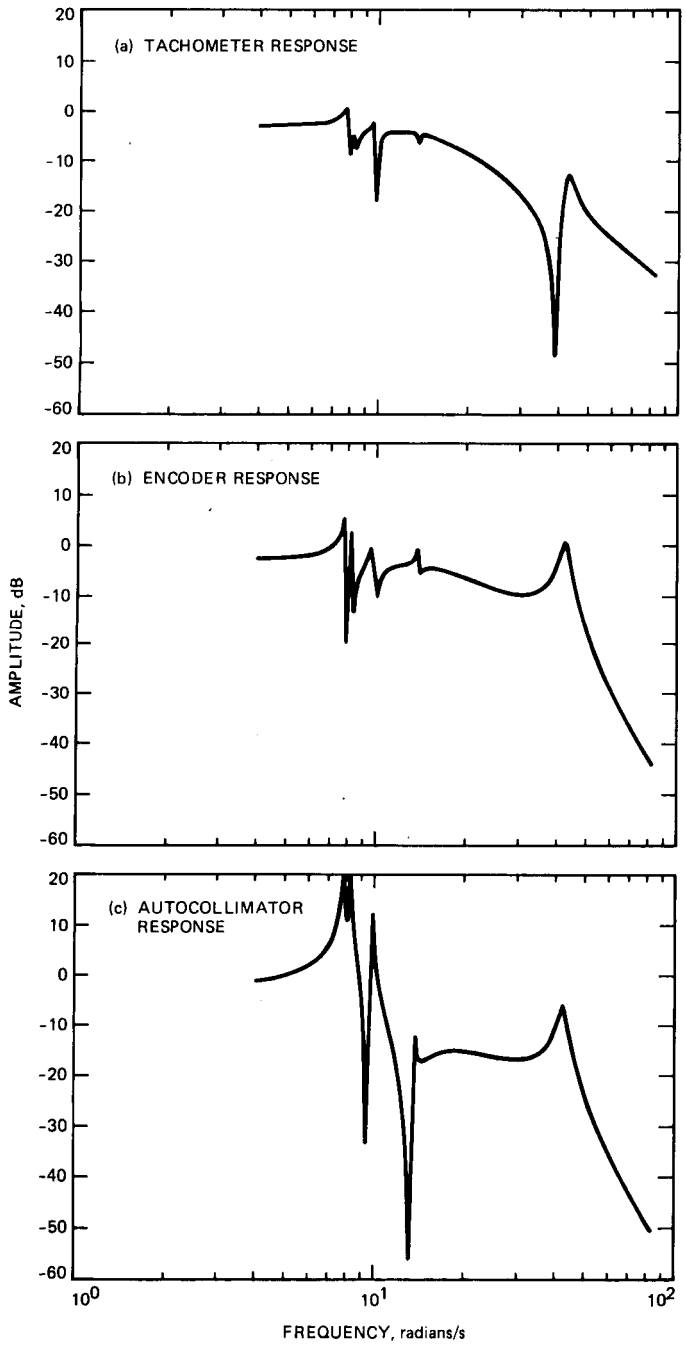


Fig. 5. Amplitude response vs. frequency for 70-m azimuth

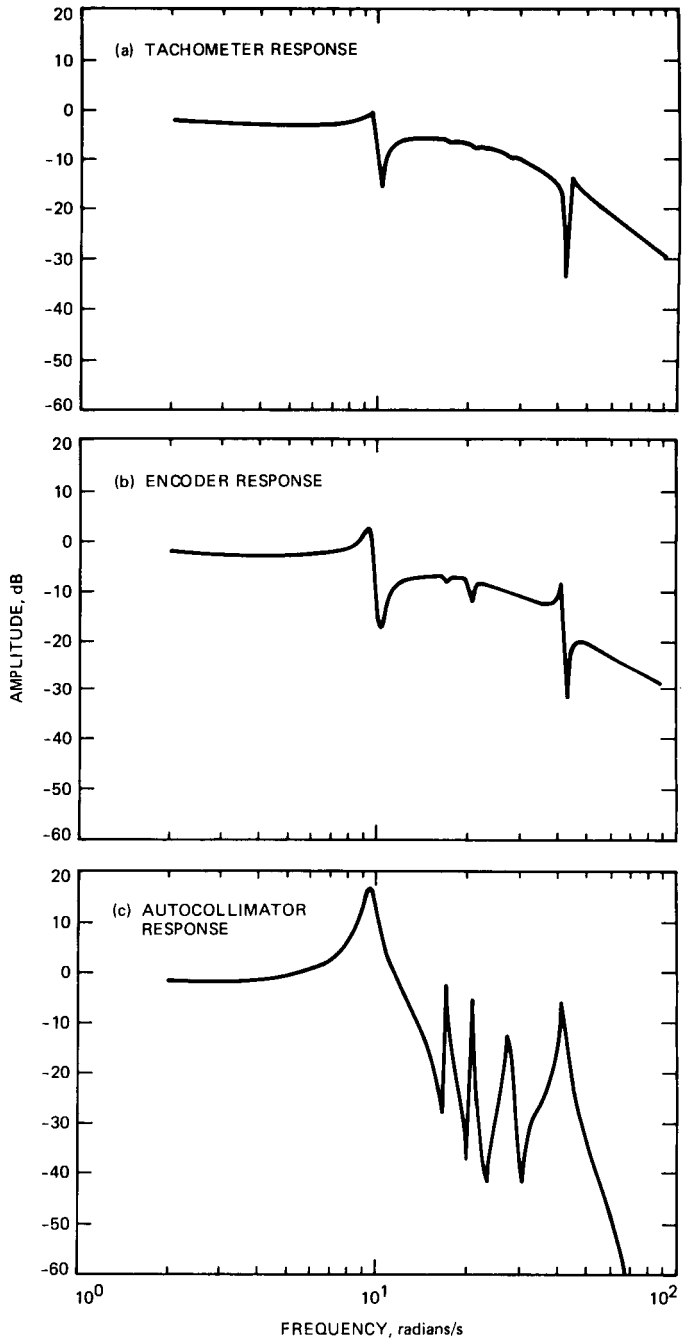


Fig. 6. Elevation response vs. frequency, 70-m antenna

X-Band Uplink Ground Systems Development: Part II

C. E. Johns
Radio Frequency and Microwave Subsystems Section

The prototype X-band exciter testing has been completed. Stability and single-sideband phase noise measurements have been made on the X-band exciter signal (7.145–7.235 GHz) and on the coherent X- and S-band receiver test signals (8.4–8.5 GHz and 2.29–2.3 GHz) generated within the exciter equipment. Outputs are well within error budgets.

I. Introduction

Previous performance tests made on the prototype exciter have been reported [1], but at that writing, stability measurements of the exciter and the X- and S-band test signals had not been performed. In [1], a brief description of the exciter mechanization and the method of generating its reference signals was given, and some measurement data on the exciter phase-correcting and command verification loops was presented. Since that report, exciter signal stability measurements have been made, and the test data is presented here.

II. Exciter Output Functions

A. Exciter Stability

The exciter stability measurements ($\Delta F/F$) were made at JPL using the maser test facility. A simplified diagram of the test setup is shown in Fig. 1. The exciter signal was set to 7.2 GHz + 1 Hz, and phase was compared with a stable 7.2-GHz signal generated from a hydrogen maser. The 1-Hz difference signal at the phase comparator output was fed to the Allan variance computer. In addition, the 100 MHz from the maser was supplied to the 100-MHz input ports on the exciter and to the reference input on the exciter synthesizer (Dana).

The results of the exciter stability measurements are shown in Fig. 2. The stability at the 1000-s integration period is about 2.5×10^{-16}, which is nearly an order of magnitude better than the stability budget allotted to the exciter alone (1.7^{-15}) for the overall X-band uplink system.

Also measured was the single-sideband phase noise density ($S\phi$), measured in a 1-Hz bandwidth, for frequencies from 1 Hz to 20 kHz. The results are shown in Fig. 3. Shown on the graph is the measurement of the upper limit of the test system noise floor. For the X-band uplink system, the specified maximum noise density at the 1-Hz offset from the carrier is –50 dBc, and at a 1000-Hz offset, the specified maximum is –70 dBc. The measured data shown in Fig. 3 indicates that the exciter signal phase noise is far below the specified limits.

B. X-Band Test Signal Stability

The stability of the 8.4-GHz coherent receiver test signal is shown in Fig. 4. The data shows the stability at the

1000-s integration period to be about 2.5×10^{-16}, the same as the exciter signal. This is to be expected, since the X-band test signal (F_{x-x}) comprises about 85 percent of the exciter output signal (F_x) and 15 percent of the coherent translator reference signal (131 $F_x/749$). The stability specification for the X-band test signal in the X-band uplink system is $\leqslant 2.75 \times 10^{-15}$ at 1000 s.

The single-sideband phase noise measurement of the X-band test signal is shown in Fig. 5. Also included is the upper limit of the test system noise floor. No specified limits exist for the X-band test signal phase noise in the X-band uplink system.

C. S-Band Test Signal Stability

The frequency stability of the 2.3-GHz coherent test signal is shown in Fig. 6. At the 1000-s integration period, the stability of the signal is about 7×10^{-16}. This is also to be expected, since the S-band signal comprises only about 68 percent of the exciter signal and 32 percent of the coherent translator signal (509 $F_x/749$) that is generated by a 8144/749 frequency shifter module followed by a X5 frequency multiplier. Like the X-band test signal, the specified S-band stability is $\leqslant 2.75 \times 10^{-15}$ at 1000 s for the X-band uplink system. The measured single-sideband phase noise of the S-band signal and the upper limit of the test system noise floor are shown in Fig. 7. As with the X-band test signal, there are no specified phase noise limits for the S-band test signal.

III. Conclusions

The measured exciter output and the X- and S-band test signal stabilities show that the signals are well within their specified error budget for the Deep Space Network X-band uplink project. Also, stability measurements made on the original X-band exciter at DSS-13 have shown that the long-term stability, including the instability of the uncompensated 43-MHz cable from the control room to the antenna equipment, is within 2.5×10^{-15} [2].

References

[1] C. E. Johns, "Block III X-Band Receiver-Exciter," *TDA Progress Report 42-89*, vol. January–March 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 83–93, May 15, 1987.

[2] T. Y. Otoshi and M. M. Franco, "DSS-13 Stability Tests," *TDA Progress Report 42–89*, vol. January–March 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 1-20, May 15, 1987.

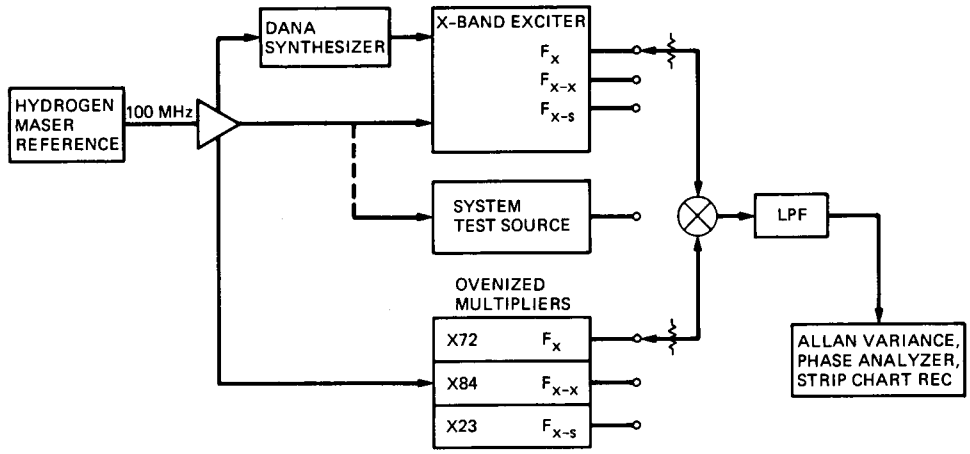


Fig. 1. Exciter test setup block diagram

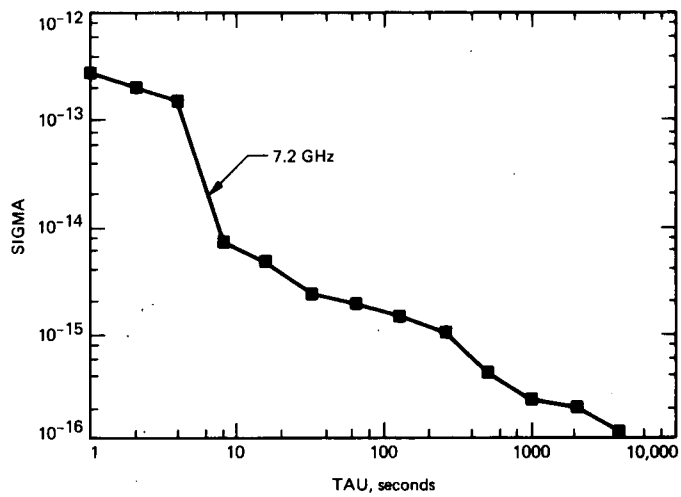


Fig. 2. Exciter output stability

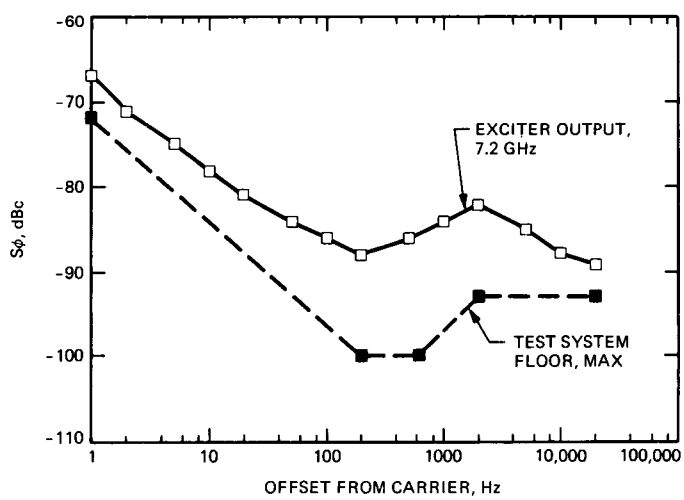


Fig. 3. Exciter phase noise

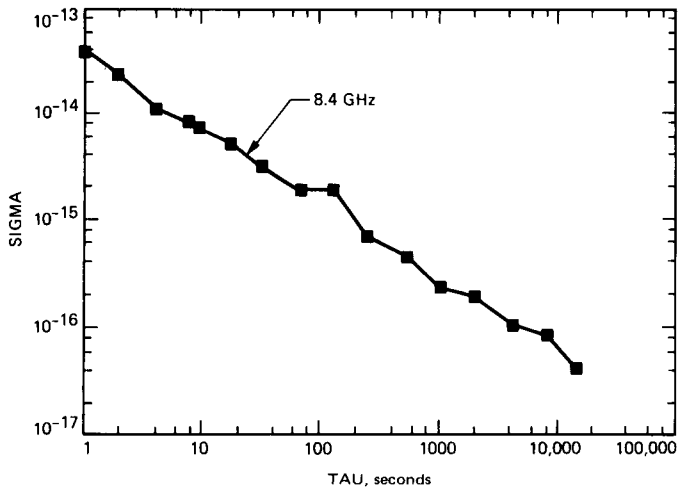


Fig. 4. X-band test signal stability

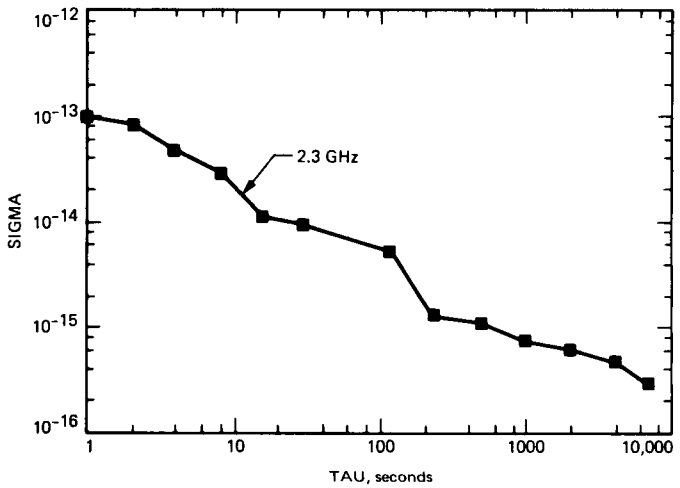


Fig. 6. S-band test signal stability

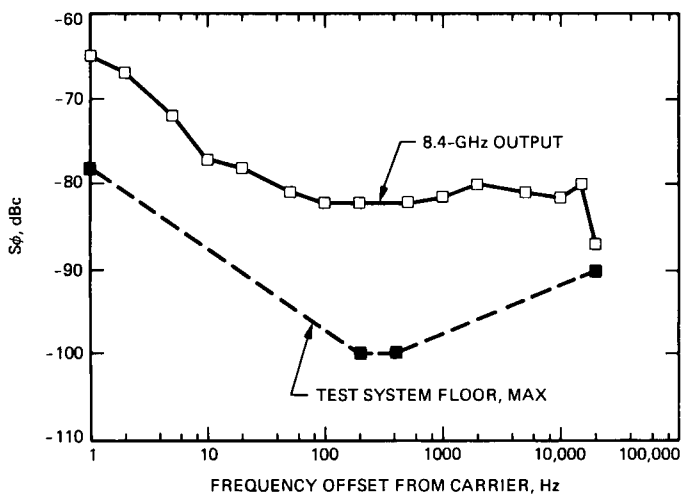


Fig. 5. X-band test signal phase noise

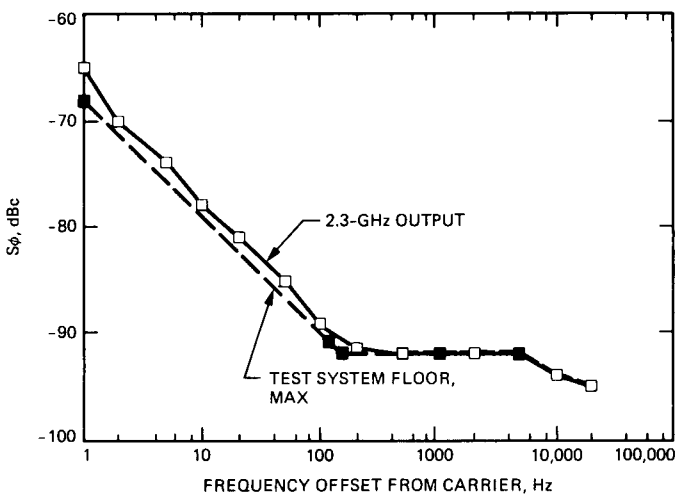


Fig. 7. S-band test signal phase noise

Analysis of the ICE Combiner for Multiple Antenna Arraying

C. Foster and M. Marina
Radio Frequency and Microwave Subsystems Section

The passage of the International Cometary Explorer (ICE) through the tail of comet Giacobini–Zinner took place on September 11, 1985, at approximately 11:04 GMT. The signal-to-noise ratio of the data received from the ICE spacecraft during the comet encounter was improved by arraying the 64-m antenna channels A and B (RCP and LCP) with the two 34-m antennas. Specially designed combiners were built to combine the signals received by the three antennas at the different DSN sites to ensure that the spacecraft's weak signal was received. Although the ICE spacecraft was built with a 5-W transmitter and with a small antenna designed to provide data from no farther than 1 million miles, these combiners provided enough signal margin during the encounter to receive the ICE transmitted data from within the tail of the comet Giacobini–Zinner, 44 million miles from Earth.

I. Introduction

ICE telemetry link analysis showed that reception from the ICE at the Giacobini–Zinner encounter at the desired 1024-bps data rate would be made feasible by combining the two downlinks (channels A and B) received by the DSN 64-m antenna [1]. The addition of two 34-m antennas to this array would provide an additional margin of 0.8 dB to the received signal and would reduce effects caused by weather, etc. (see Table 1).

Five combiners were built for the DSN to support the capability to array together the telemetry output of the 34-m antennas at DSS-12 and DSS-15 and the 64-m antenna RCP and LCP channels. The combiners were multi-input-port (two, three, or four port), single- or triple-output-port devices to allow operational flexibility for ICE operations and testing (Fig. 1).

II. The Combiners

The function of the combiners is to accept baseband signals from the telemetry phase detector's output; to weight the signals appropriately in accordance with their respective SNRs to provide the maximum signal-to-noise ratio at the output; to sum the signals; and to provide the combined output signal to prime and backup telemetry data streams. The block diagram in Fig. 2 shows how this was accomplished. Baseband signals were passed through variable attenuators which were used to set the weighting factors to provide the maximum signal-to-noise ratio at the output of the summing amplifier. Current amplifiers were used to provide the combined output signal for prime and backup telemetry data streams.

Table 1 lists the relevant parameters for the four-input-port combiner for the ICE–GZ encounter.

III. ICE Encounter Support

Two combiners were operated at each DSN complex during the encounter, the prime unit combining all available signals and a backup combining only the two strong signals from the 64-m antenna (Fig. 2).

Baseband receiver telemetry phase detector outputs were patched directly into an input port, and the output of the combiners was patched directly to a telemetry string consisting of a Subcarrier Demodulator Assembly (SDA), a Symbol Synchronizer Assembly (SSA), and a Telemetry Processor Assembly (TPA).

The combiner required external monitoring to validate its performance. (Separate telemetry strings provided assessments of proper performance.) The SSA measured signal-to-noise ratios, and the sequential decoder provided a symbol error rate statistic. Thus, when array testing was under way, all available telemetry strings were used to measure configuration setup conditions and long-term performance.

Baseband signals were combined using optimum combining coefficients $\{\alpha_k\}$ so that the output SNR was the sum of the input SNRs. Appendix A describes the expression used for combining coefficients $\{\alpha_k\}$.

For the ICE–GZ encounter, the combining coefficients were normalized so that the total output power from the combiners was approximately the same as the baseband power levels ($\simeq$6 dBm), to be within the acceptable range for the SDAs:

$$\{\alpha_i\} = \alpha_i k; \qquad k = \frac{1}{\Sigma\alpha_i}$$

The arrival time difference for signals at the SPC-10 (DSS-14/15/12) combiners has two main components. One component is the transport delay from the receiver input on the antenna to the input port of the combiner. The transport delay for DSS-14 and DSS-15 is approximately 1.3 μs and can be ignored. The transport delay for DSS-14 and DSS-12 is approximately 55 μs, since the stations are 16.5 km apart. Because the ICE combiner does not compensate for this delay, the result is 0.2 dB loss in SNR when a signal arrives at each antenna at the same time.

The second component is the geometric time-of-arrival difference for the baseline between the DSS-14 and DSS-12 antennas. For the ICE–GZ encounter, DSS-12 will contribute about 0.3 dB to the net symbol signal-to-noise ratio (SSNR) from spacecraft rise to meridian crossing, but its contributions will deteriorate to nearly 0.1 dB by spacecraft set. This expectation was supported by observation during the test pass on July 26, 1985, at a bit rate of 1024 bps.

For the time near comet encounter, the arrival time delay for DSS-12 to DSS-14 followed approximately the following:

Spacecraft Position	T (DSS-12/DSS-14)
Rise	6 μs
Meridian	12 μs
Set	−35 μs

Appendix B derives the expected SNR degradation due to arrival time dephasing during the Giacobini–Zinner comet encounter.

IV. Conclusions

Figure 3 shows the observed performance at Goldstone using the three- and four-input-port combiners for this time interval. The actual improvement in signal-to-noise ratio realized by the ICE resistive combiners can be estimated from the plots of the symbol error rate (SER) and the SSNR performed on DOY 185 (1985) using the three-input-port combiner to combine DSS-14 (channels A and B) with DSS-12. These plots are shown in Fig. 4.

At this time, the ICE spacecraft was operating at a bit rate of 512 bps. The sum of the DSS-14 channels A and B, with SSNRs of 1.3 dB and −0.2 dB, yields a theoretical combined SSNR of 3.68 dB, while the observed SSNR was 3.6 dB. Thus, the actual improvement due to antenna arraying using the ICE resistive combiner was within 0.1 dB of the calculated prediction. The combiner is currently being used to support Pioneer and ICE spacecraft.

Reference

[1] N. A. Fanelli, L. Efron, and R. J. Muellerschoen, "ICE Second Halley Radial: TDA Mission Support and DSN Operations," *TDA Progress Report 42-87*, vol. July–September 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 285–290, November 15, 1986.

Table 1. Symbol definitions and values

Symbol	Description	Value
R_1	SSNR at DSS-14 (RCP)	0 dB
R_2	SSNR at DSS-14 (LCP)	−1 dB
R_3	SSNR at DSS-12	−7 dB
R_4	SSNR at DSS-15	−8 dB
r_s	Symbol rate	2048 sps
B_{rf}	Baseband bandwidth	730 kHz
α_1	DSS-14 (RCP)	0.52
α_2	DSS-14 (LCP)	0.48
α_3	DSS-12	0.16
α_4	DSS-15	0.14
$R_1 + R_2$	Combined SSNR (DSS-14 only)	2.55 dB
$R_1 + R_2 + R_3 + R_4$	Combined SSNR (DSS-14, 12, 15)	3.34 dB

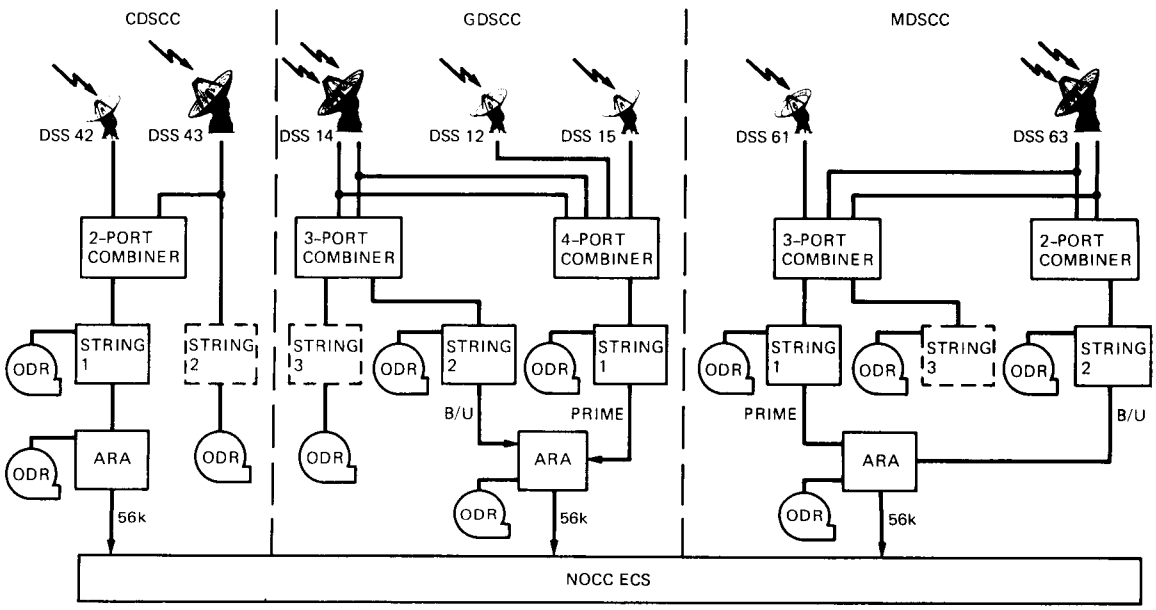


Fig. 1. ICE encounter support configuration

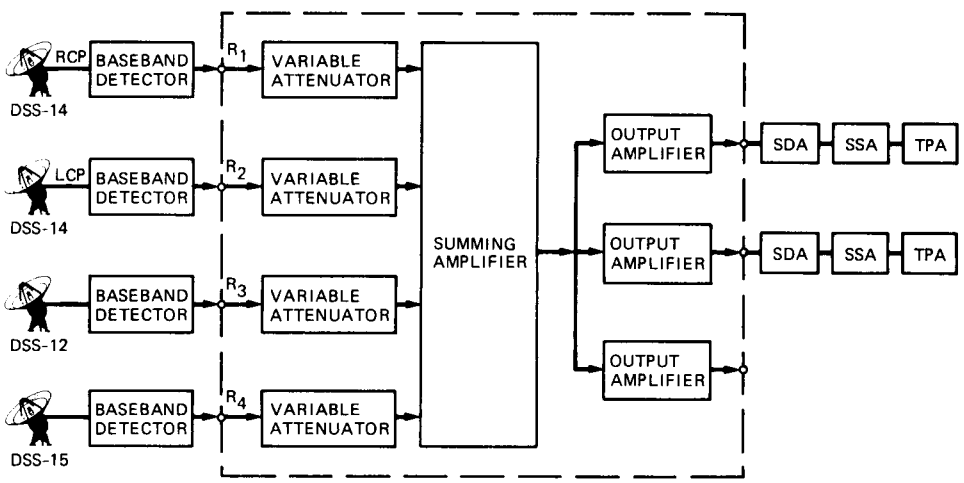


Fig. 2. Four-input-port combiner block diagram

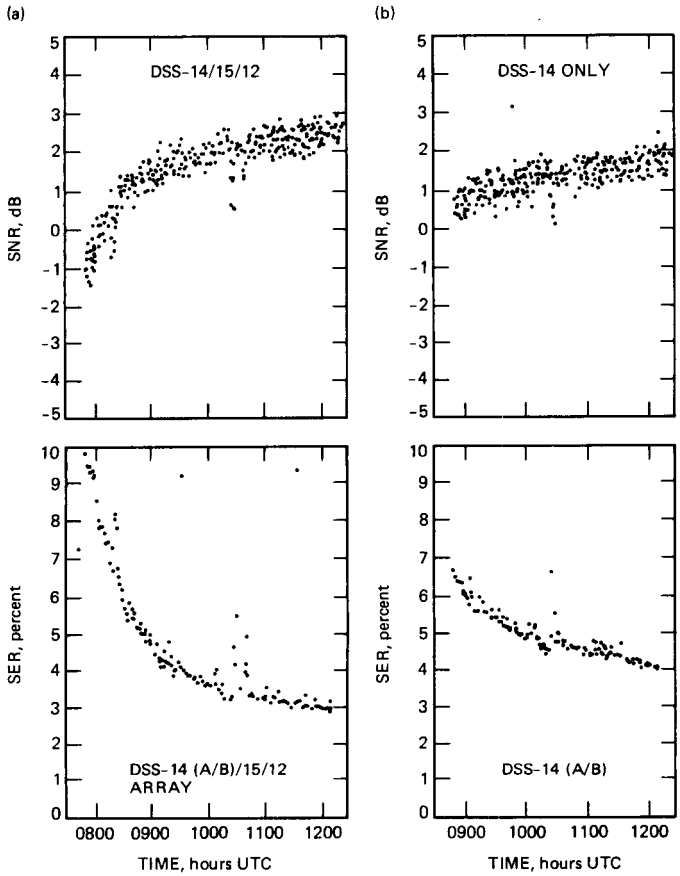


Fig. 3. Observed ICE link performance at encounter (comet tail crossing = 1104 Z on DOY 254): (a) four-port combiner; (b) three-port combiner

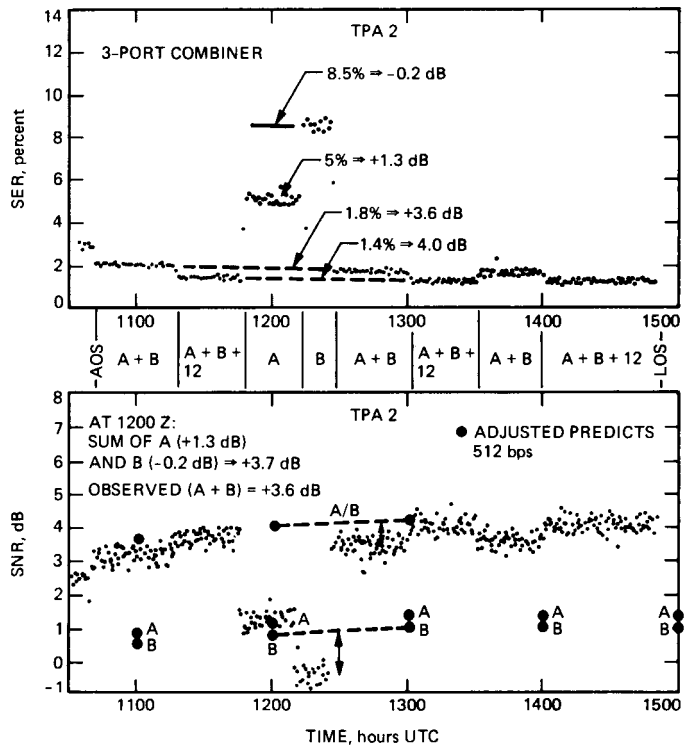


Fig. 4. Performance of Goldstone at 512 bps on 1985 DOY 185

Appendix A
Arraying Methodology

If we have n Gaussian voltage sources with means u_K and standard deviation σ_K, we define the signal-to-noise power ratio R_K as

$$R_K = \frac{u_K^2}{2\sigma_K^2} \qquad K = 1, \ldots, n \tag{A1}$$

where u_K is the mean integrated symbol voltage, σ_K is the rms integrated symbol noise voltage, and the channel total power is

$$P_K = u_K^2 + \sigma_K^2 \tag{A2}$$

These Gaussian sources model the integrated symbol voltage distribution from each of n telemetry chains with independent noise

$$\rho_K(x) = \frac{1}{\sigma_K\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x \pm u_K}{\sigma_K}\right)^2\right\} \tag{A3}$$

signal voltage density. Now, if we combine the signals with n combining coefficients $\{\alpha_K\}$, we get, for signal voltages that add coherently:

$$u_A = \sum_{K=1}^{n} \alpha_K u_K \tag{A4}$$

for noise powers that add incoherently:

$$\sigma_A^2 = \sum_{K=1}^{n} \alpha_K^2 \sigma_K^2 \tag{A5}$$

Then

$$R_A = \frac{u_A^2}{2\sigma_A^2} \tag{A6}$$

is the combined array SNR and

$$P_A = u_A^2 + \sigma_A^2 \tag{A7}$$

is the combined output channel total power.

The optimum combining ratios $\{\alpha_K\}$ to maximize R_A with respect to the $\{\alpha_K\}$ are found by taking $\partial_{RA}/\partial_{\alpha K} = \phi$.

Thus:

$$\frac{u_A}{\sigma_A^2}\frac{\partial_{uA}}{\partial\alpha_K} - \frac{u_A^2}{2\sigma_A^2}\frac{1}{\sigma_A^2}\frac{\partial\sigma_A^2}{\partial\alpha_K} = \phi \tag{A8}$$

And using (A2) and (A3), we get

$$\alpha_K = \left(\frac{u_K}{u_A}\right)\left(\frac{\sigma_A^2}{\sigma_K^2}\right)\left(\frac{u_A}{u_K}\right)\left(\frac{R_K}{R_A}\right) \tag{A9}$$

From (A1) and (A2), we get

$$u_K = \sqrt{\frac{2P_K R_K}{1+2R_K}} \tag{A10}$$

From (A5) and (A6), we get

$$u_A = \sqrt{\frac{2P_A R_A}{1+2R_A}} \tag{A11}$$

Thus:

$$\alpha_K = \frac{R_K}{R_A}\sqrt{\frac{P_A R_A (1+2R_K)}{P_K R_K (1+2R_A)}} = \sqrt{\frac{P_A R_K (1+2R_K)}{P_K R_A (1+2R_A)}} \tag{A12}$$

In the case where $P_K = P_A$, we get

$$\alpha_K = \sqrt{\frac{R_K\,(1+2\,R_K)}{R_A\,(1+2\,R_A)}} \tag{A13}$$

Using these values for $\{\alpha_K\}$, we have

$$R_A = \frac{u_A^2}{2\sigma_A^2} = \frac{\left(\sum_{K=1}^{n} \alpha_K\right)^2}{2\sum_{K=1}^{n} \alpha_K^2\,\sigma_K^2}$$

$$= \sum_{K=1}^{n} \frac{u_K^2}{2\sigma_K^2} = \sum_{K=1}^{n} R_K \tag{A14}$$

So for the above choice of $\{\alpha_K\}$, the output SNR is the sum of the input SNRs. Note that the combined SNR R_A is the same if all $\{\alpha_K$'s$\}$ are multiplied by some constant value.

In particular, we can normalize so the largest α_K is equal to unity $\bar{\alpha}_K = \alpha_K/\alpha_R$; $\alpha_R = \max\ \{\alpha_K\}$.

Appendix B

Degradation Due to Dephasing

During the Giacobini-Zinner comet encounter, the ICE spacecraft was operating at a bit rate (r_B) of 1024 bps with a rate 1/2 code and biphase modulation format.

The symbol period $T_s = 1/r_s$ where r_s is the symbol rate where $r_s = 2r_b$. The Manchester transition period T_x is

$$T_x = \frac{1}{2} T_s = \frac{1}{4r_B} \tag{B1}$$

where $r_B = 1/T_B$.

Now let ρ_{xs} be the symbol transition probability and let ρ_{xm} be the Manchester transition probability.

Thus,

$$\rho_{xm} = 1 - \frac{\rho_{xs}}{2} \tag{B2}$$

so the average number of transitions per second N_x is

$$N_x = \frac{\rho_{xm}}{T_x} = 4\,\rho_{xm}\,r_B \tag{B3}$$

for *PN* symbols ($\rho_{xs} \simeq 0.5$) $N_x \simeq 3r_B$.

With no dephasing, the combined output signal $u_A(0)$ becomes:

$$u_A(0) = \sum_K u_K \tag{B4}$$

and the in-phase power

$$u_A^2(0) = \sum_K \sum_1 u_K\,u_1 \tag{B5}$$

Now let any of the Kth channels be dephased by Ψ_K seconds (see Fig. B1). Then the combined output signal voltage $u_A(\Psi_K)$ will be:

$$u_A(\Psi_K) = \sum_K u_K\,(1 - 2N_x\,|\Psi_K|) \quad \text{where } K = 1, \ldots, n \tag{B6}$$

while the output noise power σ_A^2 is unaffected.

$$u_A(\Psi_K) = u_A(0) - 2N_x \sum_K u_K\,|\Psi_K|$$

$$= u_A(0)\left[1 - 2N_x \sum_K \frac{u_K}{u_A(0)}\,|\Psi_K|\right] \tag{B7}$$

Thus:

$$u_A(\Psi_K) = u_A(0)\,\eta_A(\Psi_K) \tag{B8}$$

where

$$\eta_A(\Psi_K) = \left[1 - 2N_x \sum_K \frac{u_K}{u_A(0)}\,|\Psi_K|\right] \tag{B9}$$

Using $ln\,(1 - x) \simeq -x$ when $x << 1$, then the amplitude loss in nats of the combined output signal becomes

$$ln\,\eta_A(\Psi_K) \simeq 2N_x \sum_K \frac{u_K}{u_A(0)}\,|\Psi_K| \tag{B10}$$

and the power loss

$$ln\,\eta\rho\,(\Psi_K) = 2\,ln\,\eta_A(\Psi_K) \simeq 4N_x \sum_K \frac{u_K}{u_A(0)}\,|\Psi_K| \tag{B11}$$

In the case where $2N_x \Psi_K \alpha_K u_K / u_A << 1$ is not true, then we have:

$$u_A(\Psi_K) = u_A(0)\left(1 - 2N_x \Psi_K \frac{\alpha_K u_K}{u_K}\right)$$

$$= u_A(0)\left(1 - 2N_x \Psi_K \frac{u_K^2 \sigma_A^2}{u_A^2 \sigma_K^2}\right) \quad \text{(B12)}$$

and from Eq. B9 we have

$$u_A(\Psi_K) = u_A(0)\left(1 - 2N_x \Psi_K \frac{R_K}{R_A}\right)^2 = u_A^2(0) \cdot X \quad \text{(B13)}$$

Thus, the SNR degradation is given by X.

Example: An array configuration with no dephasing and a bit rate (r_B) of 1024 bps has a theoretical combined output SSNR of R_A = 3.35 dB. Dephasing one channel (whose contribution is approximately 0.45 dB to the net SSNR) by 50 microseconds will degrade the SSNR by 0.2 dB.

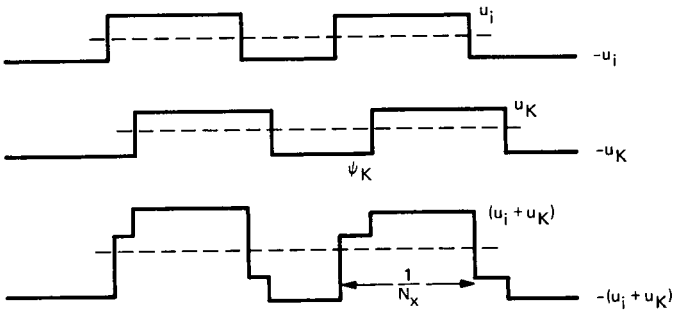


Fig. B1. Two combined channels dephased by Ψ_K

The 8.4-GHz Low-Noise Maser Pump Source Assembly

R. Cardenas
Radio Frequency and Microwave Subsystems Section

Improved pump source assemblies and new 8.4-GHz low-noise traveling-wave masers (TWMs) were installed at the same time at Deep Space Stations 14 and 43 as part of the Mark IVA DSCC Antenna Microwave Subsystems upgrade. The pump source assemblies are part of the new 8.4-GHz TWMs, which are identified as Block IIA Low-Noise TWMs. Improved reliability of the pump source assemblies was required to meet stress analysis criteria.

I. Introduction

A redesign of the 8.4-GHz low-noise pump source assembly and its controller, described previously in [1] and [2], was made in accordance with stress analysis criteria set by the Microwave Electronics Group in accordance with parts derating requirements set forth in JPL Specification ZPP-2061-PPL [1] The design was also a part of the Mark IVA DSCC Antenna Microwave Subsystems upgrade.

The evaluation of the reliability history of the pump source assembly resulted in a redesign of the modulator and protective circuit, which is a subassembly of the pump source assembly. Figure 1 shows the pump source assembly that is used on the Block IIA TWMs, and Fig. 2 shows the modulator and protective circuit subassembly. In the modulator and protective circuit subassembly, the improvements and changes were made to the components on the two printed wiring boards that comprise the circuit portion of this subassembly.

The protective circuit portion of the modulator and protective circuit subassembly limits the voltage supplied and protects against a reverse polarity voltage to the varactor tuning element and the Gunn bias voltage, thereby protecting the two Gunn-effect oscillators from overvoltage and wrong polarity. The dual Gunn-effect oscillators in the pump source assembly operate at two different frequencies: 19 GHz and 24 GHz. The reason for protecting the Gunn-effect oscillators is to improve the mean time between failures of the maser system. Also, the cost of each Gunn-effect oscillator is high, and the procurement-and-repair time is anywhere from 9 to 18 months.

II. History

The first boards used in the modulator and protective circuit subassembly were hard-wired boards. These boards were based on the pump sources used on R&D masers. The hard-wired boards were reconfigured into a printed wiring board format to facilitate reproducibility; the existing components were used in the same manner as the previous design. Reports from the field pointed out that these particular boards had a reliability problem. The blocking diode, CR1, tended to burn out as a result of inadequate power rating (see Fig. 3), causing failure of the pump source assembly.

During the redesign of the pump source assembly, it became evident that a number of components were not operating

[1] "Preferred Parts List, Reliable Electronic Components," JPL Specification ZPP-2061-PPL (internal document), Jet Propulsion Laboratory, Pasadena, California, July 1982.

within their maximum rating levels. Another problem encountered with the redesign of the modulator and protective circuit subassembly was that some components were no longer readily available from the commercial vendors.

The modulator and protective circuit subassembly's printed wiring boards were redesigned so that the new boards would be interchangeable. Any future upgrades can be made simply by either swapping out or modifying the existing modulator and protective circuit subassembly in the pump source assembly.

III. Design Changes

The design changes described below were made to provide additional protection of the Gunn-effect oscillators from any damage and to upgrade the pump source assembly to meet the stress analysis reliability requirements set by the Microwave Electronics Group in accordance with the requirements of JPL Specification ZPP-2061-PPL.[1]

A stress analysis was carried out, taking into account a number of design constraints that were imposed by the subassemblies involved. In accordance with DSN STD 00001,[2] the TWMs require a type 1 environment that is air conditioned in a manner similar to the tricone environment (25°C to 35°C), thus imposing tolerance requirements on pump source components. It was also necessary to ensure that any changes made to the modulator and protective circuit boards would not require any changes to the existing chassis size.

The stress analysis was made under the worst case tolerance of each of the design parameters, and an evaluation was made by determining the maximum stress on each component compared to the derating factors listed in JPL Specification ZPP-2061-PPL.[1] Data was taken under simulated worst case conditions, and the redesigned modulator and protective circuit subassembly met or exceeded the requirements.

The following changes were made to components of the modulator and protective circuit subassembly (see Fig. 4):

(1) The tolerance value of the key resistors was changed from ±5 percent to ±1 percent. The ±1 percent tolerance was chosen to ensure that under the worst case analysis, the failure protection predictions would be repeatable, as well as to protect the overall circuitry. Resistors were changed to make the circuit performance repeatable. Another resistor was changed from a lower to a higher wattage value to comply with the stress rating.

(2) Transistor Q1 was changed from a lower to a higher breakdown voltage. The collector emitter breakdown voltage of the old transistor was 40 volts, and the applied power supply voltage in some cases may have been greater than this. Reported failures in the field support this finding.

(3) Blocking diode CR1, which protects the Gunn bias circuit from reverse polarity, was changed. This diode, which is no longer made, was the component that failed most often in the field. It was determined under laboratory conditions that the old blocking diode did not meet the environmental requirements and was overheating and burning out. The new diode has a faster-acting reverse recovery time of 25 nanoseconds. A black anodized aluminum block heat sink was added to dissipate the heat from this blocking diode. The heat sink operates at a maximum temperature of 89°C, which is well below the manufacturer's specification for the blocking diode (175°C maximum).

(4) Zener diode VR1 was changed from a higher (-51 volts) to a lower (-45 volts) clamping circuit voltage, limiting the voltage applied to the varactor tuning element in the Gunn-effect oscillator. A second zener diode, VR2, was added in parallel with VR1 to provide redundant protection. Figure 5 is a chart of the output performance of VR1 and VR2.

(5) Removable jumpers JTB1 and JTB2 were added (see Fig. 6). These jumpers allow individual testing of the zener diodes.

(6) The existing printed circuit board was modified to accept the above changes and additions (Fig. 6).

(7) No change was required of the overvoltage protector devices A2 and A3, located in the pump source assembly (see Fig. 1). The overvoltage protector device prevents excessive power dissipation in the Gunn diode and is set at some voltage above the normal operating Gunn bias voltage for each of the Gunn-effect oscillators.

The following changes were also made to the maser pump controller assembly (see Figs. 7 and 8):

(1) Overvoltage protectors OVPPS1 and OVPPS2 were added. The protectors were required (Fig. 8) because the subassembly was not capable of supplying enough protection to the pump source assembly to prevent any wide open supply voltage from being applied to the modulator circuit or the varactor diode of the Gunn-

[2]"DSN Standard Design Requirements, DSIF/GCF/NCS Equipment," DSN Standard 00001 (internal document), Jet Propulsion Laboratory, Pasadena, California, February 1986.

effect oscillator. One overvoltage protector is set across the modulator power supply, which is capable of an output voltage of 120 volts if the output voltage adjusting pot opens. The device is set at 45 volts, which is the maximum voltage required by the modulator circuit. The second overvoltage protector is set across the tuning power supply in the pump control assembly and is set at 70 volts to protect the varactor diode in the event that the adjusting potentiometer for the supply opens. These changes served to further protect the redesigned modulator, the protective circuit subassembly, and the Gunn-effect oscillators.

(2) A zener diode VR3, rated at 82 volts, 50 watts, was added. This zener diode is connected across the tuning power supply output. A heat sink was also added to dissipate the 10 watts of power when this component is activated. This zener diode is a backup for OVPPS2. In the event that OVPPS2 does not fire, this 82-volt diode will limit the voltage so that the current through zener diodes VR1 and VR2 will not rise to a value in excess of its continuous rating. With this protection, the fault may exist indefinitely without damaging the Gunn-effect oscillators or any other component.

The overvoltage protectors and the zener diode make up the overvoltage protection subassembly (shown in Fig. 8).

IV. Conclusions

The redesigned pump source assembly and the pump source controller assembly have met and exceeded the stress analysis criteria defined in JPL Specification ZPP-2061-PPL,[1] as demonstrated by lab stress testing and by successful performance of the equipment in the Deep Space Network. Before any changes were made, approximately nine modulator and protective circuit subassemblies (each containing two boards) had failed and were retrofitted with the new, redesigned modulator and protective circuit subassembly. The retrofitting task began around July 1983. Unfortunately, an exact count was not kept of the number of subassemblies that have been replaced in the Deep Space Network. From June 1986 to September 1987, however, none of the new modulator and protective circuit subassemblies and maser pump source controller subassemblies have failed.

It is interesting to note that Gunn-effect oscillator failure has not been directly connected to modulator and protective circuit board failure, but since the new boards have been installed, it has been noted by the Cognizant Operations Engineer that fewer Gunn-effect oscillators have been replaced because of failure. According to a previous MTBF analysis initiated by R. Stevens and C. P. Wiggins in October 1983, 16 percent of the total failures in the TWM/CCR system were the result of miscellaneous electronic failures. The pump source assembly was a fraction of that percentage.

Acknowledgments

The improvements made to the modulator and protective circuit subassembly described in this article are the product of much hard work by A. P. Wagner of the JPL Electric Power Systems Section. Special thanks are due to him for his initial reliability analysis of the modulator and protective circuitry. The author also wishes to thank D. L. Trowbridge for his continuous support and guidance in this effort, as well as the members of the Microwave Electronics Group under the supervision of S. M. Petty.

References

[1] D. L. Trowbridge, "X-Band, Low-Noise, Traveling-Wave Maser," *DSN Progress Report 42-60*, vol. September–October 1980, Jet Propulsion Laboratory, Pasadena, California, pp. 126–131, December 15, 1980.

[2] D. L. Trowbridge and J. Loreman, "S-Band Ultralow Noise Traveling-Wave Maser," *Deep Space Network Progress Report 42-53*, Jet Propulsion Laboratory, Pasadena, California, pp. 148–154, October 15, 1979.

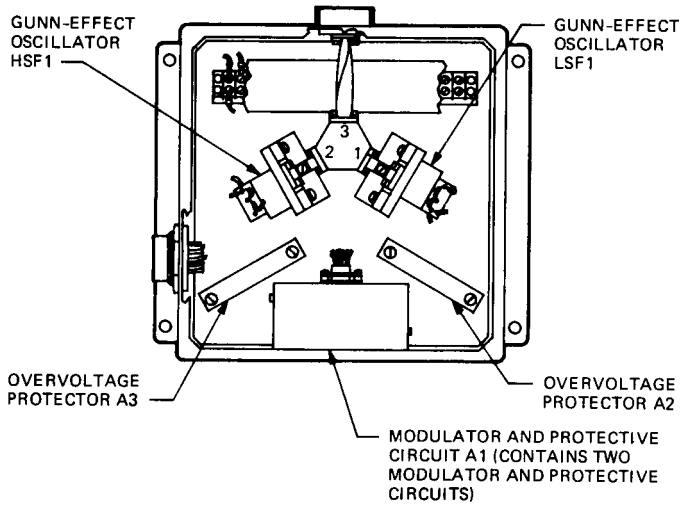


Fig. 1. The 8.4-GHz low-noise pump source assembly

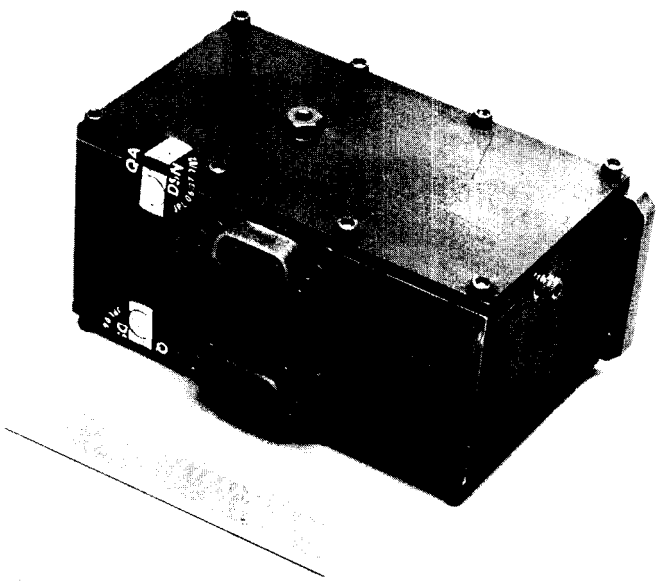


Fig. 2. Modulator and protective circuit subassembly

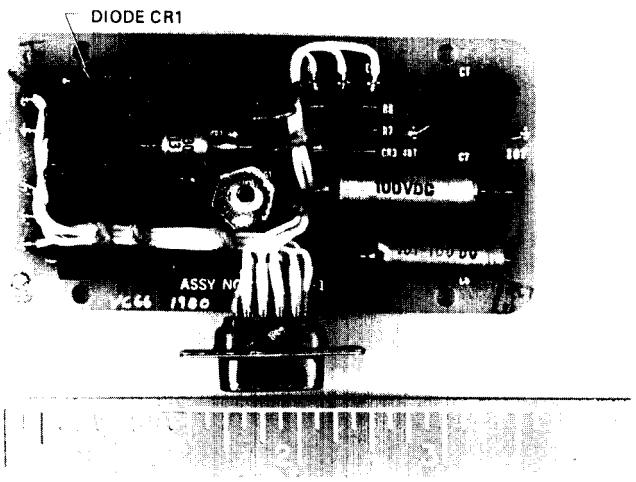


Fig. 3. Modulator and protective circuit printed wiring assembly with burned-out diode CR1 (rating: 45 ± 1 percent)

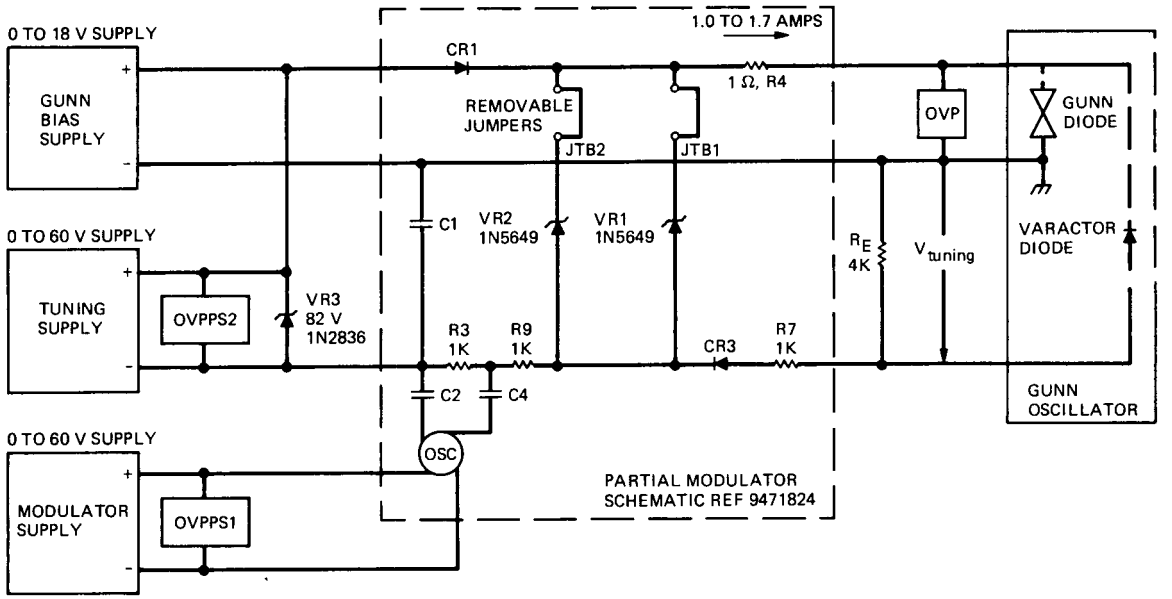


Fig. 4. Simplified schematic of overall pump source assembly

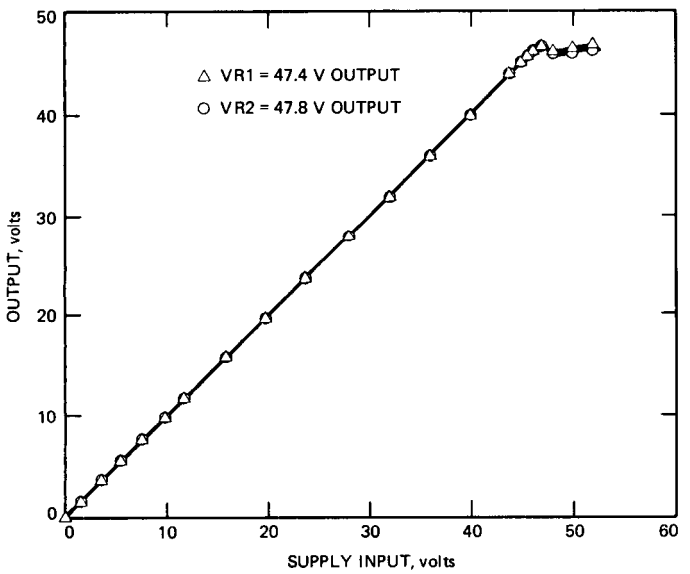


Fig. 5. Clamping voltage test data for modulator and protective circuit subassembly zener diodes VR1 and VR2

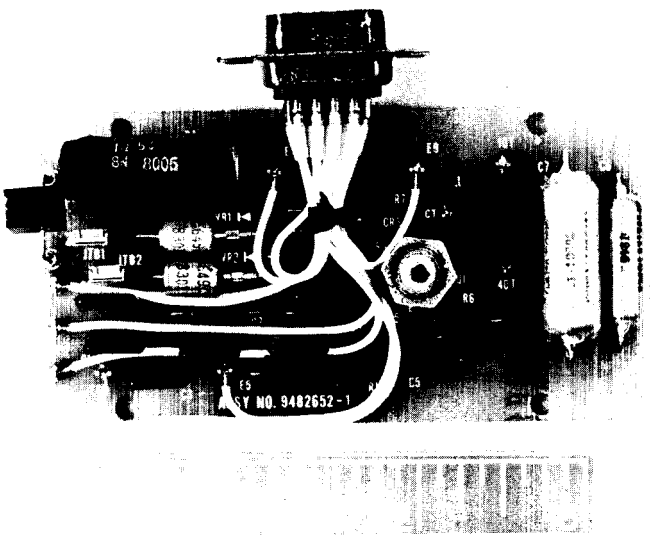


Fig. 6. Modulator and protective circuit printed wiring assembly

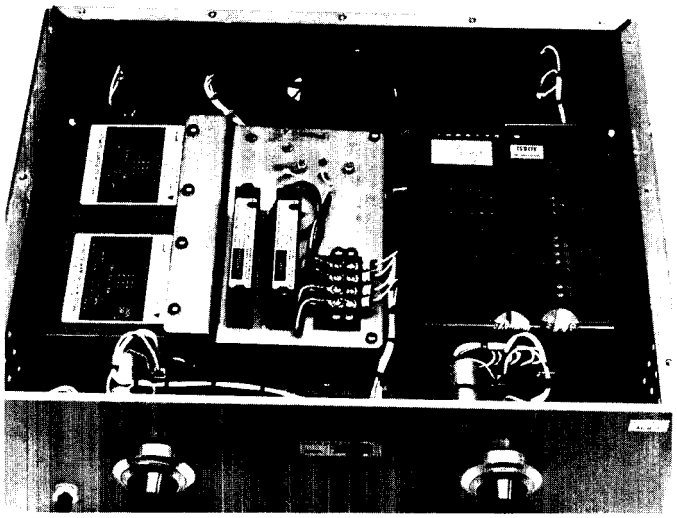


Fig. 7. Pump control with overvoltage protection subassembly installed

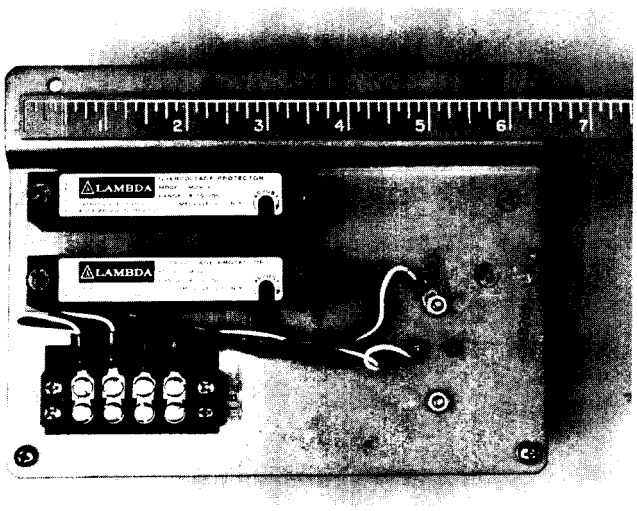


Fig. 8. Overvoltage protection subassembly

A Modern Control Theory Based Algorithm for Control of the NASA/JPL 70-Meter Antenna Axis Servos

R. E. Hill
Ground Antennas and Facilities Engineering Section

A digital computer-based state variable controller has been designed and applied to the 70-m antenna axis servos. The general equations and structure of the algorithm and provisions for alternate position error feedback modes to accommodate intertarget slew, encoder referenced tracking, and precision tracking modes are described. Development of the discrete time domain control model and computation of estimator and control gain parameters based on closed loop pole placement criteria are discussed. The new algorithm has been successfully implemented and tested in the 70-m antenna at Deep Space Station (DSS) 63 in Spain.

I. Introduction

Servo design studies utilizing dynamic models of the new 70-m antenna structures identified many changes in control dynamics which result from the antenna upgrade from 64 m to 70 m [1]. These studies indicated that the 64-m rate loop hardware and software required only minor parameter value modifications when upgraded to the 70-m system. The 70-m Antenna Servo Controller employs a software pointing control algorithm similar to that used in the previous 64-m antenna configuration. This article describes the new servo control algorithm and the analytic servo design methods employed in deriving the control coefficient values for the new 70-m antenna.

II. Control Algorithm Description

The position control loop is closed in the Antenna Servo Control (ASC) computer by a digital, state variable controller based on modern feedback control techniques. It employs feedback of the hardware system state variables to achieve the required closed loop system performance. A linear estimator, or observer, provides a measure of those states which are not instrumented. The state variable feedback controller differs from the classical feedback controller, which relied on the use of cascade networks (implemented in either hardware or software) to provide the necessary servo compensation.

The state variable controller has several advantages: (1) its compensation techniques are not limited to the use of physically realizable networks; (2) the state variable approach facilitates discrete time domain design methods, resulting in efficient utilization of the control computer; (3) the state estimator provides an accurate, low-noise measure of rate and acceleration; (4) the estimator provides a powerful method of encoder data error detection and correction which permits uninterrupted operation during brief intervals when the encoder data are unusable; and (5) the state variable method

employs a standardized matrix representation of the system which is convenient for digital computer processing. The closed loop state variable controller has the linear system properties of a classical linear feedback controller, and its performance is related to bandwidth and linear error coefficients.

The operation of the control algorithm is illustrated in the block diagram of Fig. 1, which depicts a single axis of the two-axis Azimuth/Elevation system. The block labeled "Antenna Hardware" represents the control electronics, the rate loop, and the output gear ratio. Its output, **X**, is a vector whose elements represent the individual system states. The hardware control input, U, corresponds to the rate command input to the hardware generated by the control algorithm. Antenna position is sensed by a bull gear referenced angle encoder and also by an optical autocollimator. These devices are represented by two additional blocks, both having the input state vector **X**. The autocollimator detects antenna position error relative to a Precision Instrument Mount or Master Equatorial (ME) which serves as a precision tracking position reference and is controlled in celestial hour angle and declination coordinates. The sensing axes of the autocollimator are precisely aligned with respect to the geometric axis of the antenna primary reflector, thus providing more precise pointing than is available from axis mounted angle encoders.

The representative equation of the "Antenna Hardware" block is the generalized difference equation relating the hardware state vector at the discrete times of the computer sampling to the state vector's previous value and to the control input, U. The discrete transition matrix, Φ, and the input vector, Γ, describe the dynamic behavior of the physical system. The angle encoder and the autocollimator are represented by vectors $\mathbf{H_E}$ and $\mathbf{H_A}$, which operate on the state vector, **X**, and produce the scalar azimuth (or elevation) encoder angle and autocollimator angle error.

The software estimator computes the antenna state variables for subsequent processing and feedback to the hardware control input. The estimator is essentially a dynamic simulation of the physical antenna and has the same input and difference equation as the hardware. Under ideal conditions, the estimator output state vector, **E**, is identical to the antenna state vector, **X**. Errors in the estimated state vector, **E**, arise from modeling errors and from uncompensated disturbances. These errors are corrected by a comparison of the estimated with the true values of the encoder angle, and feedback of the estimator error, Y_E, with a feedback gain factor, **L**.

The feedback of the estimator error introduces a feedback loop around the estimator with a dynamic response governed by the amount of correction (gain) in the loop. By proper selection of the estimator correction coefficients, **L**, the speed of correction of erroneous estimates (and also the speed of response to transient invalid encoder data) can be adjusted. The coefficients are therefore designed to reach a compromise between rapid estimator error corrections and encoder data noise filtering.

A logical test on the size of the estimator error provides a powerful method of detecting gross errors in encoder reading. The omission of the estimator correction when such errors are detected is equivalent to substituting the estimated value for the real encoder value and thus provides an optimal filter of gross errors. The limits on such a test must be sufficiently wide to permit recovery after an interval of rejecting erroneous data. Estimator errors due to modeling and disturbances will tend to accumulate during periods of invalid data and could prevent the acceptance of valid data if the limits are too restrictive. A dynamic limit test, with time and rate dependent limits calculated to accommodate cumulative error, has been considered but not yet implemented.

The servo feedback loop is closed by multiplying the state estimate, **E**, by a feedback gain, **K**. Because the estimate, **E**, is approximately equal to the hardware output, **X**, the result is an effective feedback loop around the physical hardware. The dynamic properties of the closed servo loop are hence controlled by the value of **K**. The input to the software rate/acceleration limiter is

$$U = N \cdot R - \mathbf{K} \cdot \mathbf{E} \tag{1}$$

where R is the input command, N is the input gain constant, and $\mathbf{K} \cdot \mathbf{E}$ is the scalar product which when expanded yields

$$U = N \cdot R - K_1 \cdot E_1 - K_2 \cdot E_2 - K_3 \cdot E_3 - \cdots - K_6 \cdot E_6 \tag{2}$$

The input gain, N, controls the overall gain of the servo and is assigned different values according to the mode of operation.

The individual elements of gain vector **K** correspond to integral error, position, rate, acceleration, and other gains, and thus determine the closed loop stiffness and dynamic response of the servo. The K values are assigned to achieve the desired linear system performance. The method of computation of the K values is described in detail in Section V.

III. Modes of Operation

A. Computer Small Error Mode

In the Computer Small Error Mode the servo is configured as described above in Section II to provide linear type II posi-

tion servo performance. This mode is employed when position control utilizing feedback from the axis encoders is required. The rate command is computed according to Eq. (2) and the input gain, N, is set equal to K_2 to provide unity position gain. Automatic transfer to the "Large Error" mode occurs any time conditions which would result in prolonged control saturation occur. A simple test initiates the transfer any time the calculated control input, U, exceeds 1.5 times the software rate limit.

B. Precision Mode

In the Precision Mode, the Az/El servo tracks the position of the Master Equatorial (ME), as contrasted to the Computer Small Error Mode where positioning to a specific encoder angle is required. The angular difference between the antenna and the ME is detected by a two-axis optical autocollimator mounted on the antenna. A single axis of the autocollimator is depicted in Fig. 1 as a hardware block with inputs from both the (Az/El) hardware and the Master Equatorial. The autocollimator output is filtered by a 2-pole low pass filter in the hardware. In software, the azimuth error signal is multiplied by the approximate secant of the elevation angle to correct for the elevation-dependent geometry of the autocollimator reflected light path.

Precision Mode position control is accomplished through an alternate branch in the software which evaluates control and estimator equations modified to use autocollimator derived position error in place of the position command, R, and the estimated encoder angle E_2. The modified estimator equations compute integral error based on the filtered autocollimator error, as contrasted to the encoder vs command angle error which is used in the computer mode. In the Precision Mode the input to the software rate/acceleration limiter becomes

$$U = -K_1 \cdot E_1 - K_2 \cdot \epsilon_P - K_3 \cdot E_3 - \cdots - K_6 \cdot E_6 \qquad (3)$$

The modified control and estimator equations provide type II control using the autocollimator error for position and integral control with damping provided by the estimator. A distinct advantage of this configuration over true error-only control is the continuity of the estimator equations between the computer and precision modes. Since the integral error estimate, E_1, changes relatively slowly, there is no abrupt change in the estimator output resulting from mode to mode transitions. Mode transfer settling times are thus minimized.

The software employs a four-pole minimum Integral Time Absolute Error (ITAE) low pass digital filter to restrict the error signal bandwidth and eliminate signal components resulting from deflections of the alidade structure. The filter bandwidth of 2.86 Hz was adjusted empirically on the 64-m antenna to provide the best damping of the alidade structure and to minimize the effects of autocollimator noise.

C. Large Error Mode

To prevent saturation of the type II servo loop, a third mode of operation is provided with software controlled mode selection. A Large Error Mode of operation is provided to accommodate intertarget slew motions and any other transient conditions which would saturate the type II servo and cause large excursions of position error.

In the Large Error Mode the software configuration is altered slightly from the computer mode configuration to provide slew rate control with smooth rate transitions. The estimator functions the same as in the Small Error Mode except that all of the estimator elements are not utilized in the control. The control gain, **K**, is chosen to produce a slow response rate servo with the bandwidth selected to limit the peak acceleration in response to a maximum rate step input. This type of control was chosen to minimize acceleration transients which would excite oscillations of the antenna structure. In this mode the first two elements of **K**, which correspond to integral error and position feedback gain, are set to zero. The remaining elements are selected to achieve the desired closed loop servo bandwidth. In the Large Error Mode, the command input, R, is a rate signal value computed to reposition the antenna to the desired angle in minimum time. The input gain, N, is set equal to a function of K_3, K_4, K_5, K_6 to provide unity rate servo gain.

The software transfers from Large Error to Computer Small Error control when both the rate error and position error are sufficiently small to permit unsaturated type II servo performance. The limits for transfer are 0.05 degree/s rate error and 0.03 degree position error. Upon initial entry into Small Error control, the integral angular position error estimate, E_1, is initialized to a value proportional to the estimated rate. This helps to minimize the error settling time. The mode control logic permits entry into Precision Mode only from the Computer Small Error Mode.

IV. Control Algorithm Sequencing and Timing

The computations of the discrete system matrices and of the feedback gains are based on a specific (50 ms) sample interval and negligible time delay between each encoder input and the corresponding rate command output. These conditions are satisfied by the use of timed interrupt driven software with lower priority interrupts masked during execution of time critical control computations. Time skew errors are minimized by assigning highest priority to a 100 pps interrupt which ini-

tiates encoder read and control computation at regular 50 ms intervals. Computing time effects are minimized by consecutive sequencing of the Azimuth and Elevation axis functions with the Elevation read and compute operations beginning 20 ms behind those for Azimuth. This 20 ms offset allows sufficient time for completion of the Azimuth computations.

The estimator is initialized prior to antenna brake release by equating the angular position estimate to the encoder reading and by setting all other state estimates to zero. On the transition from the Large Error to Small Error mode the integral position error estimate is initialized to a value proportional to the estimated rate. The estimator is not reinitialized on transitions to or from the Precision Mode.

V. Computation of the Control Algorithm Coefficients

A. Linear System Matrices F and G

The simplest linear dynamic model of the 70-m Az/El antenna assumes a rigid body structure with the moment of inertia dominated by that of the inertia wheels attached to the motor shafts. The compressibility of the hydraulic oil and its piping combined with the inertia result in a complex conjugate pole pair. Two real, open-loop poles corresponding to the rate loop compensation networks are also modeled. With the two integrations operating on the angular rate and position to produce position and integral error, the order of the dynamic system for control becomes six. Because estimator feedback is based on encoder angle, the integral error is an unobservable state and is excluded from the estimator design process. It is introduced later in the control gain design process.

The simplified structure model described above was used in the design of the MK IV 64-m antenna servos. However, because of the significant dynamics changes resulting from the structure additions for the 70-m antenna, a more comprehensive structure model was needed. The new model [1] includes gear reducer stiffness, three tipping structure modes, and two alidade modes in Elevation. The resulting closed rate loop model is 10th order for Azimuth and 14th order for Elevation. The low frequency closed rate loop poles of these models differ considerably from those computed from the simplified rigid body structure model even though the rigid body inertia includes the static inertia of the structure. This difference results from the finite compliance of the gear reducers and of the structure. The models for both the 70-m and 64-m structures indicate the Elevation axis compliance is dominated by the alidade.

Because of the uncertain degradation of robustness associated with errors in modeling the structure, a decision was made to reduce the new model to the sixth order in the form of the simplified model described earlier. The underlying assumption, that system robustness resulting from the use of a rigid body based model is superior to that from a higher degree structure model, has not been investigated. The reduction was accomplished by deleting the higher frequency pole-zero pairs associated with the structure modes. The remaining hydraulic motor and compensation network poles and zeros were then combined with the rate-to-position-to-integral-error integrations to synthesize the sixth order model.

The poles and zeros for the Azimuth and Elevation closed rate loops are listed in Table 1, and their corresponding linear system matrices are in Table 2. In the linear system matrices the value of 20 in the first row results from normalization of the integral error with respect to the sampling time interval. This produces a value of unity in the discrete transition matrix and helps to reduce estimator computation time. Further simplifications are accomplished by replacing the negligibly small elements in the first row of the transition matrix and the first element of the input matrix with zeros. The resulting coordinate skew corrupts the integral estimate by less than 12.5 microdegree-seconds in the worst case.

The input large error mode gain constants, N, are calculated from the linear system matrices for the condition of unity rate gain. The results for azimuth are expressed by

$$N = 1.0 + K_3 + 5.173\,K_5 + 10.001\,K_6 \qquad (4)$$

and for elevation

$$N = 1.0 + K_3 + 9.041\,K_5 + 17.479\,K_6 \qquad (5)$$

B. The Discrete Transition and Input Matrices

The linear system matrices **F** and **G** are transformed from the continuous time domain to the discrete (sampled data) time domain to produce the discrete transition matrix, Φ, and the discrete input matrix, Γ. A computer sampling time interval of 50 ms (20 samples/s) was selected, as it satisfies the criteria of (1) negligible pointing error resulting from sampling effects at maximum tracking rate and acceleration; (2) the rate exceeds 10 times the required closed position loop bandwidth; and (3) the rate being a convenient multiple of the DSN frequency and timing standard 100 pps signals. The third criterion derives from the necessity to synchronize the tracking position commands to the Deep Space Network clock.

Using the conversions for sampling described by a zero order hold with no delay

$$\Phi = \text{expm}\,(\mathbf{F} * T) \tag{6}$$

$$\Gamma = \Psi * T * \mathbf{G} \tag{7}$$

where expm denotes the matrix exponential function.

The Ψ matrix is obtained from the relationship

$$\mathbf{F} * T * \Psi = \Phi - \mathbf{I} \tag{8}$$

where **I** is the identity matrix.

A Fortran program was developed to evaluate Φ, Ψ, and Γ from the continuous time system matrices, **F** and **G**, and sampling time, T, using the scale and square method described by Moler and Van Loan [4]. The numerical results for ϕ and Γ are listed in Table 3.

The control gain vector, **K**, and the estimator gain **L** are calculated by the method of closed loop eigenvalue assignment. This method, also referred to as pole placement, combines the desired closed loop pole locations with the discrete system matrices Φ and Γ in the Ackerman [3] equations to produce the control feedback gain, **K**, or estimator gain, **L**.

The specified pole locations were iterated to insure satisfactory robustness and insensitivity to computational roundoff and to angle encoder quantizing. Deadbeat response (minimum settling time) is impractical because it requires excessive control effort to overcome small disturbances. The following general criteria were employed in specifying closed loop pole locations:

(1) The two lowest frequency poles are specified to achieve the desired bandwidth and settling time. For optimal error settling time (ITAE criteria) [2], these poles are complex conjugates with equal real and imaginary parts. Assigning more than two dominant poles using the ITAE criteria, the Butterworth criteria, or (presumably) a similar criterion tends to increase the transient overshoot of the closed loop system and is avoided. The bandwidth of the estimator is always at least three times that of the overall system and narrow enough to provide a level of noise filtering.

(2) Other low frequency poles may be specified to cancel open loop zeros in order to produce a flat closed loop low frequency response.

(3) High frequency poles are specified near their corresponding open loop locations to minimize the value of the resulting **K** or **L**. Estimator poles should be displaced slightly from the control poles to enhance robustness.

(4) The resulting values of **K** or **L** are reviewed for suitably large values of elements corresponding to integral error, position, and rate, and for small values of the remaining elements, and the pole specifications are adjusted accordingly. The control input, U, and the subsequent encoder position change resulting from single least significant bit changes of the encoder output are evaluated from the $\mathbf{K}\cdot\mathbf{L}$ and $\Gamma\cdot\mathbf{K}\cdot\mathbf{L}$ scalar products, respectively. In general, the estimate accuracy of those states not closely coupled to the output state tends to diminish with remoteness of coupling. Therefore, noise and wasted control effort are reduced as less control authority is assigned to those states.

Tables 4 through 6 list the values of the closed loop S plane poles used in the 70-m controller design and the resulting control gain, **K**, and estimator feedback gain, **L**. The estimator gain is computed from the 5th-order discrete system matrices to avoid the unobservable integral error. Dimensional consistency with Φ is obtained by expanding **L** to the sixth order with the addition of a zero value first element. Thus Table 5 lists six **L** coefficients but only 5 estimator poles.

VI. Summary

The control algorithm along with the parameter values described above were incorporated in the system software and tested successfully in the 70-m antenna at DSS 63 in Spain.

VII. Areas for Further Investigation

The degree of system performance improvement to be gained from the use of higher-order structure models has not yet been investigated. As discussed in Section V.A, only the zero order approximation of the structure was utilized in the present work because of robustness concerns. Comparison of the models of Table 1 with those of Table 6 and Figs. 5 and 6 of [1] shows that significant dynamics are neglected in the sixth-order model. The potential improvement of both estimator accuracy and structure damping can be expected to peak at some level of model complexity. Beyond this peak, estimator noise and system robustness are expected to degrade due to modeling errors and unmodeled nonlinearities. Further work can utilize current known values of the nonlinearities, structure dynamics test data, and estimates of modeling errors.

References

[1] R. E. Hill, "A New State Space Model for the NASA/JPL 70-Meter Antenna Servo Controls," *TDA Progress Report 42-91*, vol. July-September 1987, Jet Propulsion Laboratory, Pasadena, California, November 15, 1987.

[2] A. E. Bryson and Y. C. Ho, *Applied Optimal Control*, Washington, D.C.: Hemisphere Publishing, 1975.

[3] G. F. Franklin and J. D. Powell, *Digital Control of Dynamic Systems*, Reading, Massachusetts: Addison-Wesley, 1981.

[4] C. Moler and C. Van Loan, "Nineteen Dubious Ways to Compute the Exponential of a Matrix," *SIAM Review*, vol. 20, no. 4, pp. 801–836, October 1978.

Table 1. Poles and zeros of the closed rate loop model, 70-m antenna

Axis	Poles	Zeros
Azimuth	$0.00 \pm j\ 0.00$	
	$-1.45 \pm j\ 0.0$	$-2.2 \pm j\ 0.00$
	$-60.80 \pm j\ 0.00$	$-81.0 \pm j\ 0.00$
	$-7.81 \pm j\ 13.01$	
Elevation	$0.00 \pm j\ 0.00$	
	$-1.45 \pm j\ 0.00$	$-2.2 \pm j\ 0.00$
	$-37.70 \pm j\ 0.00$	$-81.0 \pm j\ 0.00$
	$-20.81 \pm j\ 16.44$	

Table 2. Linear system matrices for 70-m rate loops

Azimuth axis

$$\mathbf{F} = \begin{bmatrix} 0 & 20 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 15.17 & 0 & 0 \\ 0 & 0 & -15.17 & -15.62 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1.45 & 0.75 \\ 0 & 0 & 0 & 0 & 0 & -60.8 \end{bmatrix}$$

$$\mathbf{G} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 608 \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Elevation axis

$$\mathbf{F} = \begin{bmatrix} 0 & 20 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 26.52 & 0 & 0 \\ 0 & 0 & -26.52 & -41.62 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1.45 & 0.75 \\ 0 & 0 & 0 & 0 & 0 & -37.7 \end{bmatrix}$$

$$\mathbf{G} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 659 \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Table 3. Discrete system matrices, 70-m antenna (sample interval, TS = 0.0500 s)

Azimuth discrete transition matrix, Φ

0.100000E+01	0.100000E+01	0.239851E-01	0.000000E+00	0.000000E+00	0.000000E+00
0.000000E+00	0.100000E+01	0.461185E-01	0.141961E-01	0.251086E-03	0.136192E-03
0.000000E+00	0.000000E+00	0.784645E+00	0.477875E+00	0.138320E-01	0.610394E-02
0.000000E+00	0.000000E+00	-.477875E+00	0.292594E+00	0.301792E-01	0.772115E-02
0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.930066E+00	0.111487E-01
0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.478353E-01

Azimuth discrete input matrix, Γ

0.000000E+00
0.122145E-02
0.828046E-01
0.244640E+00
0.250242E+00
0.952165E+01

Elevation discrete transition matrix, Φ

0.100000E+01	0.100000E+01	0.225781E-01	0.000000E+00	0.000000E+00	0.000000E+00
0.000000E+00	0.100000E+01	0.413012E-01	0.162868E-01	0.321490E-03	0.209996E-03
0.000000E+00	0.000000E+00	0.568074E+00	0.417450E+00	0.158207E-01	0.861108E-02
0.000000E+00	0.000000E+00	-.417450E+00	-.870649E-01	0.148759E-01	0.394712E-02
0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.930066E+00	0.161014E-01
0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.151829E+00

Elevation discrete input matrix, Γ

0.000000E+00
0.212184E-02
0.138387E+00
0.213978E+00
0.350851E+00
0.148261E+02

Table 4. Estimator gain coefficients, L

Axis	L_1	L_2	L_3	L_4	L_5	L_6	S plane poles			
Azimuth	0.0000	0.7398	5.1375	−7.7241	3.2878	−.0252	−2.00 ± *j* 0.00	−8.00 ± *j* 0.00	−17.0 ± *j* 17.0	−60.8 ± *j* 0.00
Elevation	0.0000	0.5711	7.3192	−13.4937	5.1036	0.0001	−2.00 ± *j* 0.00	−8.00 ± *j* 0.00	−25.0 ± *j* 25.0	−37.7 ± *j* 0.00

Table 5. Azimuth control gain coefficients, K

Mode	K_1	K_2	K_3	K_4	K_5	K_6	S plane poles			
Computer small error	0.0156	0.6863	−.2312	0.0141	0.0586	−.0318	−2.20 ± *j* 0.00	−.50 ± *j* 0.50	−10.0 ± *j* 0.00	−12.0 ± *j* 12.0
Computer small error, alternate	0.0302	0.9863	−.2013	0.0397	0.0598	−.0307	−2.20 ± *j* 0.00	−.70 ± *j* 0.70	−10.0 ± *j* 0.00	−12.0 ± *j* 12.0
Precision	0.0071	0.3590	−.3287	−.1367	0.0448	−.0453	−2.20 ± *j* 0.00	−.45 ± *j* 0.45	−5.0 ± *j* 0.00	−12.0 ± *j* 12.0
Large error	0.0000	0.0000	0.4955	−.5467	−.0389	−.1270	−0.00 ± *j* 0.00 −8.0 ± *j* 0.00	−.00 ± *j* 0.00	−2.2 ± *j* 0.00	−2.0 ± *j* 2.0

Table 6. Elevation control gain coefficients, K

Mode	K_1	K_2	K_3	K_4	K_5	K_6	S plane poles			
Computer small error	0.0281	0.9153	0.2593	0.2987	0.0136	−.0368	−2.20 ± *j* 0.00	−.70 ± *j* 0.70	−8.00 ± *j* 0.00	−18.0 ± *j* 18.0
Computer small error, alternate	0.0564	1.3565	0.2791	0.3047	0.0153	−.0351	−2.20 ± *j* 0.00	−1.00 ± *j* 1.00	−8.00 ± *j* 0.00	−18.0 ± *j* 18.0
Computer small error, alternate	0.0310	0.8229	0.2751	0.3508	0.0048	−.0454	−2.20 ± *j* 0.00	−1.00 ± *j* 1.00	−4.00 ± *j* 0.00	−18.0 ± *j* 18.0
Computer small error, alternate	0.0145	0.6370	0.2469	0.2952	0.0125	−.0379	−2.20 ± *j* 0.00	−.50 ± *j* 0.50	−8.00 ± *j* 0.00	−18.0 ± *j* 18.0
Precision	0.0050	0.2612	0.2602	0.3707	−.0017	−.0519	−2.20 ± *j* 0.00	−.45 ± *j* 0.45	−3.00 ± *j* 0.00	−18.0 ± *j* 18.0
Large error	0.0000	0.0000	2.1903	1.1171	−.0869	−.1358	−0.00 ± *j* 0.00 −20.0 ± *j* 0.00	−.00 ± *j* 0.00	−2.20 ± *j* 0.00	−2.0 ± *j* 2.0

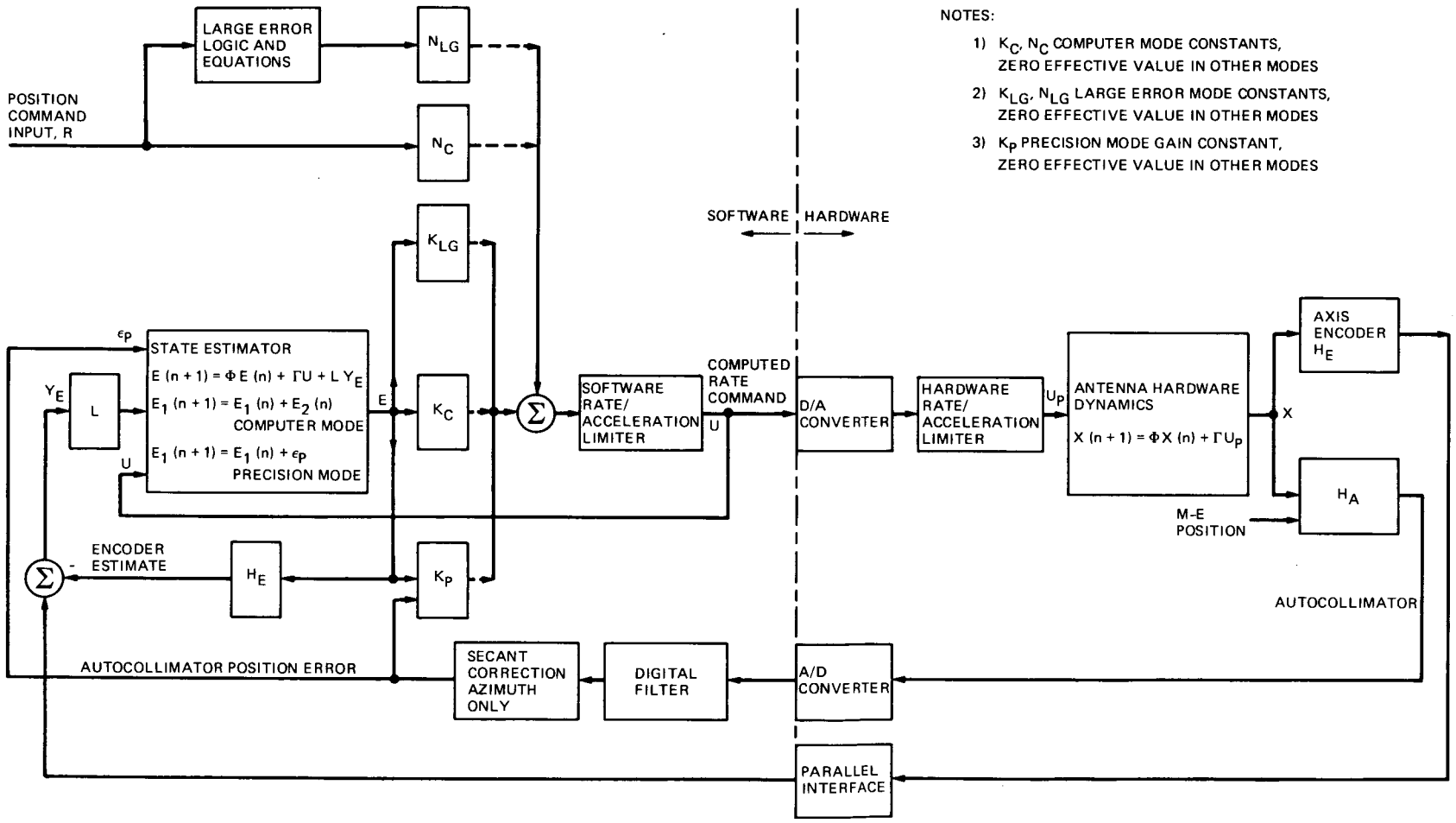


Fig. 1. Block diagram for 70-m antenna servo control algorithm

Fast Autotuning of a Hydrogen Maser by Cavity Q Modulation

G. J. Dick and T. K. Tucker
Communications Systems Research Section

A new fast autotuner for the hydrogen maser has been implemented. By modulating the cavity Q, a phase shift in the maser output signal is induced which is proportional to the cavity tuning error. This phase shift is detected and fed back to a varactor tuner to stabilize the cavity against long-term drifts. Cavity Q modulation has similarities to two other autotuning methods and significant advantages over both of them. In comparison to line Q modulation, where the frequency shift induced by a change in the atomic line Q requires a second maser for detection, the high chopping frequency allowed by cavity Q modulation gives rise to a phase shift which requires only the maser's quartz crystal "flywheel" oscillator for detection. In comparison to cavity frequency modulation, statistical noise considerations are almost identical. However, a significant advantage is the lack of phase modulation of the maser output, feedback being around the null modulation condition. Furthermore, the Q modulator is a valuable analytical tool in maser alignment. Its advantage over signal injection schemes, which give somewhat lower statistical deviation, is a lack of systematic perturbations, including independence from connecting cable lengths.

We have developed and tested a PIN-diode cavity Q modulator which gives no incidental frequency shift over a very wide range of operation. Modulated at 200 Hz, it allows variations in maser cavity frequency to be compensated with a loop gain greater than 1000. Compensation of incidental amplitude modulation of the output has also been demonstrated. Calculations show that long-term stability of $3 \times 10^{-13}/\sqrt{\tau}$ *should be achievable with typical masers.*

I. General Considerations

The hydrogen maser is the most stable frequency source generally available today for all but the shortest measuring times. For very long measuring times ($\tau > 10^4$ seconds), its performance typically deteriorates as a result of frequency drift of the high-Q resonant cavity necessary to sustain oscillation. Various schemes are used to detect this drift and then to compensate for it. This can be done periodically, as part of the setup and calibration procedure, or continuously, if allowed by the procedure. Three substantially different types of autotuning methods have been proposed and implemented.

In spin-exchange autotuning, the line width or Q of the atomic hydrogen transition itself is modulated by alternating the flow of atoms into the maser cavity between two different

rates. The line width is actually broadened by both spin-exchange and electromagnetic effects. This is because while the changing density modulates the spin-exchange interaction with other atoms, changes in the rate induce variation in the RF amplitude, and thus also in the electromagnetic transition rate bandwidth. Since the Q of the atomic line determines the efficacy of any pulling of the oscillation away from its center frequency, such a frequency error will be modulated by the change in the hydrogen flow. The accuracy of compensation is just the accuracy with which the frequency difference between the two states of operation can be determined, divided by the fractional Q modulation. Measurement of this frequency difference to the highest possible accuracy takes 1000 seconds or more and requires a second maser with equivalent stability.

An advantage of this scheme is that, to the extent that the Q modulation is due to spin exchange, the frequency offset due to this same mechanism is also eliminated, giving increased accuracy to the tuned frequency. The principal disadvantages are the need for a second maser to use as a reference and the relatively long term frequency shifts that result from the modulation, making maser output unusable during the tuning process.

In signal injection autotuning [1], [2], the frequency offset of the cavity resonator is detected as a result of the difference in cavity response at two frequencies equally spaced from the maser operating frequency but far enough from it to prevent interference with it. The injected signals are generated by offsetting the frequency of the maser output signal and, theoretically, can be much larger in amplitude, allowing an excellent signal-to-noise ratio to be obtained for the inferred cavity frequency offset. The difficulty with signal injection methods is that nearly complete carrier suppression in the injected signal is required to prevent interference with maser operation. This is because phase instabilities anywhere in the receiver and electronics used to generate and transmit the injected RF signals will modulate the frequency-pulling effect of such a carrier. The resulting frequency offset of the maser output is given by

$$\frac{\delta f}{f} = \left[\frac{\left(\frac{P_c}{P_h}\right)^{1/2}}{Q_h}\right] \cdot \sin\phi \qquad (1)$$

where P_h and P_c are the hydrogen output and injected carrier powers, respectively, Q_h is the hydrogen line Q, and ϕ is the phase difference between the injected carrier and the hydrogen signal. For a line Q of 10^9 and a required frequency stability of $\delta f/f = 10^{-15}$, some combination of overall phase stability and carrier suppression of 120 dB compared to the signal power is required by Eq. (1). Since the phase of a suppressed carrier is usually not well controlled, the entire burden is placed on its magnitude. This problem has not generally been addressed by proponents of the technique.

The injected frequencies can be applied either simultaneously or sequentially. If the signals are applied together, the requirement for carrier suppression is passed on to the injection electronics in a straightforward manner. Since the main advantage of the technique is the improvement in signal-to-noise ratio that results from large injection power, a value of

$$\frac{P_i}{P_h} = 100 \qquad (2)$$

has been proposed. From Eq. (1), and for line Q and stability as above, the required carrier suppression is 140 dB.

For switched-frequency injection, some degree of carrier suppression results from the modulation index of ~10 implied by the operating conditions specified in [2]. However, this is overcome by the larger injection power which justifies the method. At its worst, with the frequency offset f_o incorrectly chosen to be an odd multiple of the modulation frequency f_m, carrier power for square-wave frequency modulation is given by

$$\frac{P_c}{P_i} = \left(\frac{2f_m}{\pi f_o}\right)^2 \qquad (3)$$

where P_i and P_c are the injected and carrier powers, giving a phase-dependent frequency variation from Eq. (1) and Eq. (2) of

$$\frac{\delta f}{f} = \frac{\left(\frac{2}{\pi}\right)\left(\frac{P_i}{P_h}\right)^{1/2}\left(\frac{f_m}{f_o}\right)}{Q_h} \qquad (4)$$

per radian of phase difference between the injected carrier and the maser oscillation. For conditions as above and a modulation index of $f_o/f_m = 11$, Eq. (4) implies a sensitivity of about $\delta f/f = 6 \times 10^{-10}$ per radian, an unacceptable value.

A much greater degree of carrier suppression *can* be accomplished by appropriately adjusting the ratio f_o/f_m. In this case, the sensitivity to frequency and duty-cycle variation and to phase variations between the switched signals will depend in detail on the ratio chosen. For example, if time spent at each frequency is chosen to be an exact multiple of the period

defined by the frequency offset ($f_o/f_m = 2N$), the sensitivity of the remaining carrier to a timing inaccuracy δt can be shown to be given by

$$\frac{P_c}{P_i} = (2 \cdot \delta t \cdot f_m)^2 \tag{5}$$

for uncontrolled phases at the two frequencies. Restating Eq. (5) in terms of duty cycle η and modulation frequency stabilities gives

$$2 \cdot \delta\eta = \frac{\delta f_m}{f_m} = \left(\frac{P_c}{P_i}\right)^{1/2} = 10^{-7} \tag{6}$$

from Eq. (1) and Eq. (2) for $Q = 10^9$ and $\delta f/f = 10^{-15}$. If the two injected frequencies could be controlled so that they were exactly in phase at the switching points, the dependence on duty cycle would be zero to the first order, a much more attractive situation. There would remain, however, a sensitivity to any AM at the switching points of the modulation cycle. These aspects indicate the complexity associated with any realistic solutions to the carrier suppression problem.

Cavity modulation autotuning [3], [4] is similar to signal injection; in both cases the electromagnetic response of the cavity to microwave signals is used to determine its frequency relative to the maser operating frequency. The difference is that instead of modulating the signal driving the cavity, some property of the cavity itself is varied. For very rapid modulations of either the cavity Q or its frequency, the output signal from the hydrogen atoms remains relatively constant, resulting in a modulation of the amplitude or phase, respectively, of the cavity output signal, even when it is perfectly tuned. If the cavity is mistuned, a complementary phase or amplitude modulation results which is proportional to the amount of mistuning. This can be detected by means of a phase-sensitive amplifier, and the signal can be used to correct the cavity frequency.

The inherent limit to performance at long averaging times for a maser stabilized in this way is determined by the phase or amplitude noise of the output signal, depending on the type of modulation used, at the modulation frequency in relation to the signal power. In this case, the autotuning power is just that available from the maser. For frequencies of interest, namely those larger than the inverse of the hydrogen response time, the noise is typically due to the follower amplifier, being identical in phase and amplitude and independent of the modulation frequency. For this reason, limits to performance are nearly identical for the two types of modulation. Furthermore, since this same source of noise dominates maser performance for short times, the performance possible from the stabilized maser can be directly related to that of the same unit at short times without stabilization. Calculation of this relationship is presented in the following section.

Systematic errors, while inherently smaller than for signal injection, determine many aspects of the design of the modulator. Because of the large modulation complementary to that being used to detect the cavity frequency deviation, any cross-modulation effects will give rise to inferred cavity frequency deviation and thus to frequency errors in the stabilized system. On the other hand, variation of the magnitude of the desired modulation causes only a change in sensitivity to frequency error. As an example, if incidental frequency modulation $\Delta_f = \delta f_c Q_c/f_c$ accompanies an intended Q-modulation $\Delta_q = \delta Q/Q_c$, a systematic change in the output frequency results which is given by

$$\frac{\delta f}{f} = \left(\frac{\Delta_f}{\Delta_q}\right) \cdot Q_h^{-1} \tag{7}$$

where Q_c is the nominal cavity Q and Q_h is the hydrogen line Q, as previously defined. Phase shifts between the modulator and cavity cause similar effects. The advantage of the modulator is that it can be constructed of a few electronic components placed directly at the maser cavity. No cable lengths need to be interposed between it and the cavity, the thermal environment is very well controlled, and the device can be designed for insensitivity to the driving signal at the designed operating points. The crucial aspects in the design of the modulator are its long-term stability and the sensitivity of incidental cross-modulation to variation in its driving signal.

To date, long-term stability measurements have been presented only for cavity frequency modulation, even though Q modulation has some substantial advantages. These include the elimination of incidental phase modulation, which is zero in the locked condition, and the availability of variable Q for calibrating and characterizing maser performance as a function of cavity Q. In a following section we present the design for such a Q modulator and results of operational tests in a hydrogen maser. The modulator uses a PIN diode as the active element and, when properly tuned, shows no incidental frequency modulation.

II. Analysis

In this section, an analysis of autotuning by cavity modulation is presented which shows that the performance of the stabilized maser can be related in a particularly simple manner to that of the same unit without stabilization for very short measuring times. This can be done because the same

additive amplifier noise limits the statistical performance in both cases. In particular, expressions are derived for the crossover time τ_c, where the $1/\tau$ performance of the unstabilized maser and the $1/\sqrt{\tau}$ performance of the stabilized maser are equal to each other. The value of τ_c for typical conditions is approximately one second. Square wave modulation is explicitly included in the treatment, since this minimizes the problem of designing out systematic errors due to cross-modulation, as will be discussed in the following section. Both Q modulation and amplitude modulation are treated, showing only small differences between them in regard to statistical properties.

The (one-sided) spectral densities of phase and amplitude fluctuation due to additive noise are equal in value and given by [5]

$$S_\phi(f) = S_{\delta\nu/\nu}(f) = \frac{kTF}{P_o} \tag{8}$$

where k is Boltzmann's constant, T the temperature, F the noise factor of the maser receiver, and P_o the output power. This power is related to the more commonly used input power from the hydrogen P_h by

$$P_o = P_h \cdot \frac{Q_e}{Q} \tag{9}$$

where Q_e is the external Q of the cavity and Q the loaded Q. The Allan variance of frequency fluctuations in the maser output due to the effect of white phase noise as shown by Eq. (8) is given by [5]

$$\sigma_y^2 = \frac{3BkTF}{8\pi^2 f_0 P_o \tau^2} \tag{10}$$

where B is the bandwidth of the measuring system and f_0 is the operating frequency. Note that this value is 3/2 times larger than the commonly used expression for the "variance" that is due to additive white noise [6] but that does not correspond to usual data-taking procedures.

The effect of this same noise on the variance of phase or fractional amplitude fluctuations is similarly given by

$$\sigma_\phi^2 = \sigma_{\delta\nu/\nu}^2 = \frac{kTF}{2P_o\tau} \tag{11}$$

This can be directly related to a necessary uncertainty in the inferred cavity frequency, due to any phase or amplitude measurement, through the slopes of phase and amplitude with respect to frequency shown in Fig. 1. The slopes indicated in the figure are the maximum in each case. For the case of phase variation it is given by

$$\frac{d_\phi}{df_c} = \frac{2Q_c}{f_0} \tag{12}$$

from which the uncertainty in cavity frequency can be derived, with

$$\sigma_{\delta f_c} = \left(\frac{f_0}{2Q_c}\right)\sigma_\phi \tag{13}$$

If the loop is closed, inferred variations in cavity frequency will cause it to be incorrectly compensated, giving variations in the operating frequency, which is pulled by the cavity mistuning. Since the pulling of the operating frequency is given by

$$\delta f_0 = \frac{\delta f_c Q_c}{Q_h} \tag{14}$$

which, together with Eq. (13), gives

$$\sigma_y = \frac{\sigma_\phi}{2Q_h} \tag{15}$$

combining with Eq. (11) gives

$$\sigma_y^2 = \frac{\sigma_\phi^2}{4Q_h^2\tau} = \frac{kTF}{8P_oQ_h^2\tau} \tag{16}$$

for the necessary variance of fractional frequency variations under closed loop conditions. The crossover between this expression, with a logarithmic slope of $-1/2$, and the unlocked maser noise given by Eq. (10) with a slope of -1, is found to be given by

$$\tau_c = \frac{3BQ_h^2}{\pi^2 f_0^2} \tag{17}$$

which for typical conditions given by B = 20 Hz, $Q_h = 10^9$, f_0 = 1.42 GHz becomes

$$\tau_c = 3.01 \text{ seconds} \tag{18}$$

Such a crossover time is shown in Fig. 2.

If the cavity were detuned so that the operating frequency lay at the point of maximum slope of the amplitude curve

shown in Fig. 1, a similar inference could be made as to necessary fluctuations in output frequency if measurement of the cavity frequency were inferred from the resulting amplitude. As in Eq. (12) we have

$$\frac{dv}{df_c} = \frac{4v_0 Q_c}{3^{3/2} f_0} \tag{19}$$

and, following an identical procedure beginning again with Eq. (11), return a value for σ_y^2 which is larger by 27/4 than that given by Eq. (16) and a value for τ_c smaller by the same factor. This gives an apparent disadvantage to the frequency-modulation technique, but one that it recovers, as is shown below.

So far, this has been a calculation in principle, since no mechanism has been included to allow a measurement of the phase of the signal from the maser cavity. Figures 3 and 4 show the phase and amplitude variations $\Delta\phi$ and ΔV which result from rapid Q and frequency modulation in the presence of an offset between the cavity frequency f_c and the operating frequency f_0. Considering the case of Q modulation explicitly, instead of Eq. (12) we write

$$\Delta\phi = \frac{2\Delta_q Q_c \Delta f_c}{f_0} \tag{20}$$

where $\Delta_q = \Delta Q_c / Q_c$ is the fractional Q modulation, and Δf_c is the frequency offset between f_0 and f_o. Combining with the pulling Eq. (14) gives, in a manner analogous to Eq. (15)

$$\sigma_y = \frac{\sigma_{\Delta\phi}}{2\Delta_q Q_h} \tag{21}$$

The variance for the phase difference $\sigma_{\Delta\phi}$ shown in Fig. 3 is also not the same as that given by Eq. (11), but it is easy to evaluate for the case of square-wave modulation. If one-half of the time is spent in each state, the value given by Eq. (11) would double, effectively taking τ to a new value $\tau/2$. The difference between two such quantities will again double the square of the variance, giving

$$\sigma_{\Delta\phi}^2 = \frac{2kTF}{P_o \tau} \tag{22}$$

which, when combined with Eq. (21), gives a value of

$$\sigma_y^2 = \frac{kTF}{2\Delta_q^2 Q_h^2 P_o \tau} \tag{23}$$

for the variance and

$$\tau_c = \frac{3B\Delta_q^2 Q_h^2}{4\pi^2 f_0^2} \tag{24}$$

for the crossover time. For $\Delta_q^2 = 2$, and the conditions as described above, a value of

$$\tau_c = 0.375 \text{ second} \tag{25}$$

is obtained.

For frequency modulation, there is an apparent value to choose for the frequency deviation; it is just that which maximizes the slope. Taking that offset (a fractional displacement of $1/[\sqrt{8Q}]$ of the cavity frequency), modulation as shown in Fig. 4 gives a signal strength of

$$\Delta v/v = \frac{8Q_c \Delta f_c}{3^{3/2} f_0} \tag{26}$$

proportional to the frequency offset Δf_c, as in Eq. (20). Again, accounting for the statistics of modulation we have

$$\sigma_{\Delta v/v}^2 = \frac{2kTF}{P_o \tau} \tag{27}$$

as in Eq. (22) and

$$\sigma_y^2 = \frac{3^3 kTF}{2^5 Q_h^2 P_o \tau} \tag{28}$$

for the variance, giving, again with Eq. (10)

$$\tau_c = \frac{4BQ_h^2}{9\pi^2 f_0^2} \tag{29}$$

for the crossover time, and

$$\tau_c = 0.447 \text{ second} \tag{30}$$

for the conditions as above, showing a slight advantage for the frequency modulation method. This would be reversed if a Q modulator could be designed which, instead of dissipating energy in the cavity, either enhanced it or transmitted it to the receiver. Variance values calculated here are somewhat higher than those estimated in [1].

The reference signal for phase measurements can be provided by a quartz crystal oscillator. As is demonstrated in Fig. 2, for times corresponding to a modulation frequency of 100 Hz the phase noise from the maser, as calculated, will be the dominant contribution to measurement uncertainty.

III. PIN Diode Modulator

Figures 5 and 6 show block and schematic diagrams of an autotuned maser using a PIN diode Q modulator. Cavity tuning was accomplished using a varicap diode external to the maser physics package. The modulator itself is placed in the vacuum space and is mounted directly on the resonant cavity itself. First tests on a test-bed maser were entirely successful, with the locked loop sustaining its operation for indefinite periods of time and for a wide variety of time constants.

In the design of the PIN-diode modulator an attempt was made to minimize as much as possible any variation of the systematic contribution of the modulator and its support equipment to the maser output frequency. To that end, it seemed necessary to use square-wave modulation; while it might be possible to make the tuning properties of the modulator insensitive to the driving conditions at the end points, it would be difficult to accomplish this throughout the range of its operation.

Figure 7 shows schematically the tuning properties of three possibly useful modulators. Of these, the curve labeled C is clearly superior, with no detuning anywhere in its range. Curve B is less desirable but still workable. It would be necessary to minimize the switching time because of detuning effects in its midrange, but equal tuning effects at its two end points mean that duty cycle sensitivity is not a problem. Curve A is clearly the worst, showing insensitivity to applied current in the highly lossy state as required but requiring careful control of the duty cycle, since frequencies at the end points are different. The zero-current state, shown at the origin, is inherently insensitive to external circuit instabilities, since the diode is in an open-circuit condition at that point. We find that, depending on its tuning, our modulator follows closely curve A or C. Its operation can be understood as follows.

If the circuit diagram for any passive device coupled to an electromagnetic resonator is redrawn in the form shown in Fig. 8, the effect of the circuit on the properties of the resonator takes a particularly simple form. In this case the loading Q_1 and frequency shift are given by

$$Q_1 = R_{\text{eff}} \cdot \frac{2\pi E_c}{(emf)^2} \tag{31}$$

and

$$\frac{\delta f}{f} = C_{\text{eff}} \cdot \frac{(emf)^2}{2E_c} \tag{32}$$

where E_c is the energy stored in the resonator and *emf* is the open circuit voltage coupled to the circuit resulting from that energy and the coupling configuration. Reduced to this form, it is apparent that frequency shifts can be seen to be due only to an effective capacitance (positive or negative), and added losses only to the effective resistance.

The equivalent circuit of the PIN modulator shown in Fig. 6 is given in Fig. 9, showing the loop inductance L_c, tuning capacitor C_t, incidental inductance L_i, and the PIN diode parameters R_p and C_p. If the capacitance C_p is a constant, the circuit as shown is sufficient to give performance C as shown in Fig. 7. This is accomplished by tuning the variable capacitor C_t so that its reactance is equal and opposite to that of the sum of the two inductances. In that case R_{eff} in Fig. 8 becomes R_p and C_{eff} becomes C_p. Since C_p is assumed constant, only a constant frequency shift results under any circumstance, and a change in R_p only affects the Q. Specifications for PIN diode parameters often show an effective parallel capacitance which has one value under back biased conditions, and another when resistive. The addition of R_t in Fig. 10 would allow compensation for such a characteristic.

We do not find any evidence of variation in C_p as the PIN diode resistance is varied by a changing current through the diode. We chose to design the modulator for "low Q" operation, with the reactances associated with L_c and C_t about equal to the loading resistance of the diode in the "on" condition. The value of this resistance is about 100 ohms. The Q of the cavity can be reduced far below its nominal "low Q" value by further reduction of R_p to 10 ohms or below. The tuning procedure is to adjust for nominally zero frequency shift in this very low Q condition where any imbalance between $L_c + L_i$ and C_t is exacerbated. Using this procedure, no tuning effects could be detected between high and low Q states of the modulator.

IV. Summary

Analysis of several types of autotuning schemes has been presented with particular attention to both statistical and systematic errors in cavity-modulation and signal injection methods. Systematic variations due to incomplete carrier suppression in signal-injection methods are found to be a very substantial difficulty, probably outweighing any statistical advantage. Statistical analysis of cavity Q- and frequency-modulation methods shows them to be essentially identical

in this regard, with limiting performance shown to be directly related to that of the unstabilized maser.

A PIN-diode Q modulator has been designed, constructed, and tested which shows no observable incidental frequency modulation. First tests on a test-bed maser were entirely successful, with the locked loop sustaining its operation for indefinite periods of time. Operation of this relatively low performance unit was not adversely affected in any way by the effects of the modulator. Further tests are under way.

Acknowledgments

We would like to acknowledge the substantial contributions to this work by R. L. Sydnor, the cooperation and support of L. Maleki and P. F. Kuhnle, and assistance with tests and data reduction by R. E. Taylor, W. A. Diener, and C. A. Greenhall.

References

[1] C. Audoin, P. Lesage, J. Viennet, and R. Barillet, "Theory of Hydrogen-Maser Auto-Tuning System Based on the Frequency or Phase Method," *IEEE Trans. Instrum. Meas.*, vol. IM-29, pp. 98–104, June 1980.

[2] C. Audoin, "Fast Cavity Auto-Tuning Systems for Hydrogen Maser," *Revue Phys. Appl.*, vol. 16, pp. 125–130, March 1981.

[3] H. E. Peters, "Design and Performance of New Hydrogen Masers Using Cavity Frequency Switching Servos," in *Proc. 38th Annual Symp. Freq. Control*, pp. 420–427, 1984.

[4] R. B. Hayes and H. T. M. Wang, "Design for a Subcompact Q-Enhanced Active Maser," in *Proc. 38th Annual Symp. Freq. Control*, pp. 80–84, 1984.

[5] "Characterization of Frequency and Phase Noise," *Recommendations and Reports of the CCIR, Fifteenth Plenary Session*, Report 580-1, p. 91, 1982.

[6] L. S. Cutler and C. L. Searle, "Some Aspects of the Theory and Measurement of Frequency Fluctuations in Frequency Standards," *Proc. IEEE*, vol. 54, pp. 136–154, February 1966.

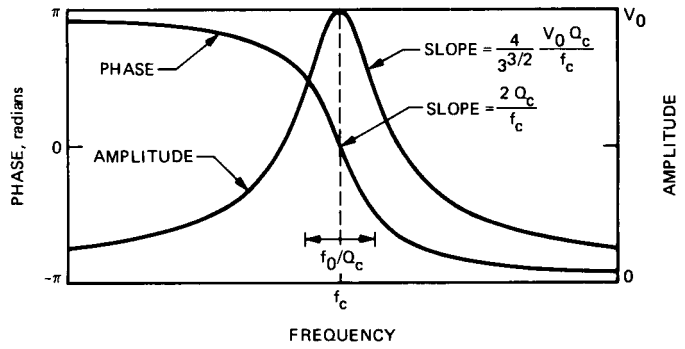


Fig. 1. Cavity frequency response

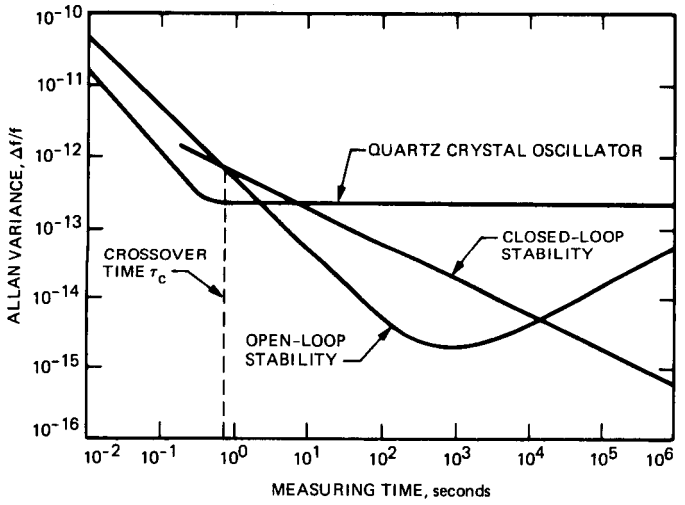


Fig. 2. Limiting performance for hydrogen maser with cavity modulation autotuning

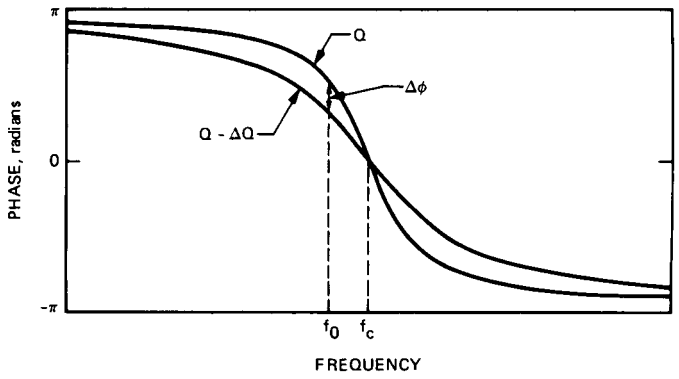


Fig. 3. Effect of Q modulation on phase response

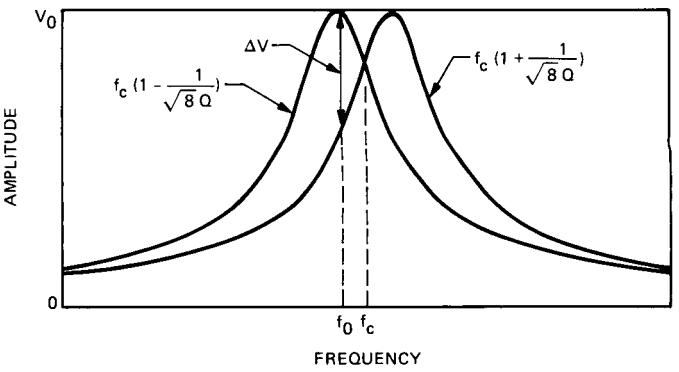


Fig. 4. Effect of frequency modulation on amplitude response

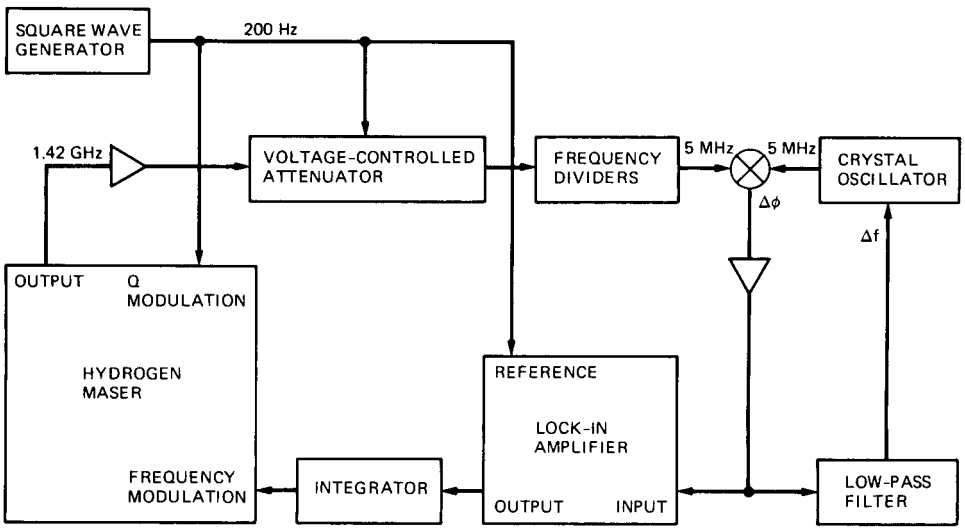


Fig. 5. Block diagram for cavity Q modulation system as tested

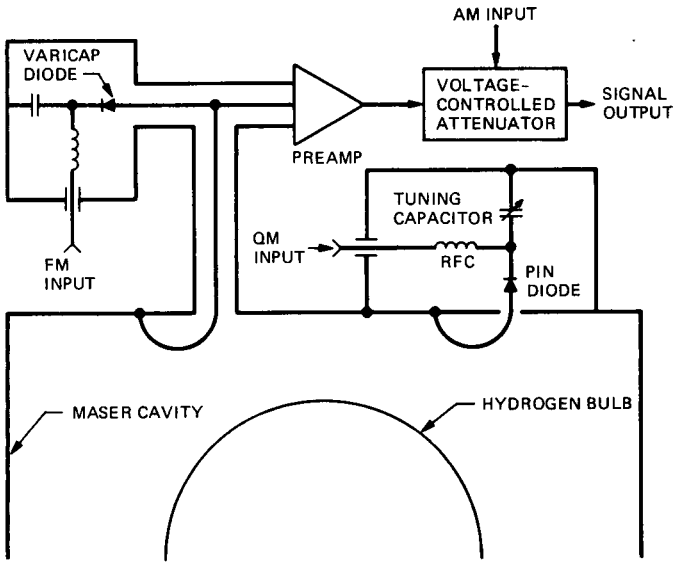


Fig. 6. Schematic diagram of Q-modulated maser

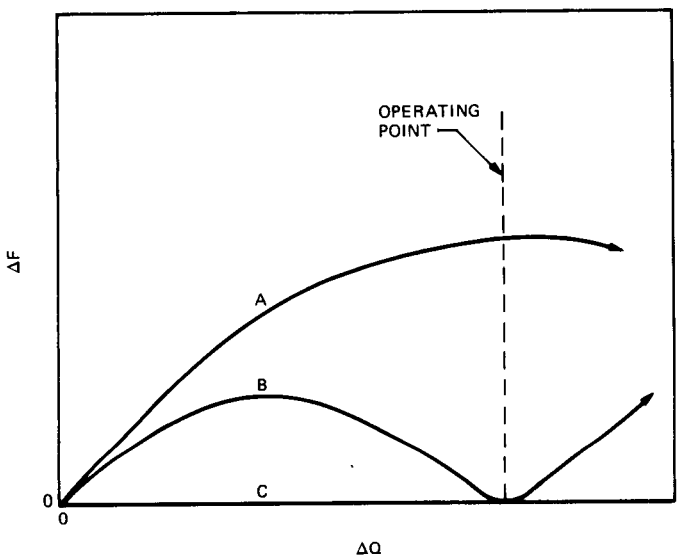


Fig. 7. Incidental frequency modulation for several usable Q modulators

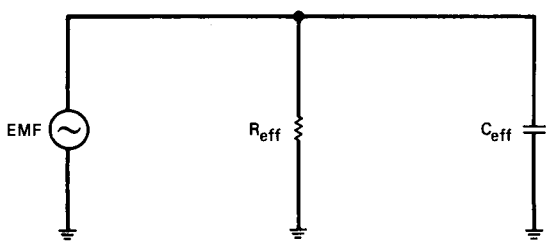


Fig. 8. Q modulator equivalent circuit

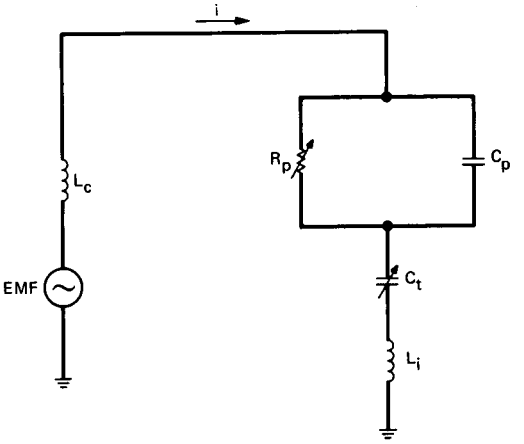


Fig. 9. Q modulator for constant C_p

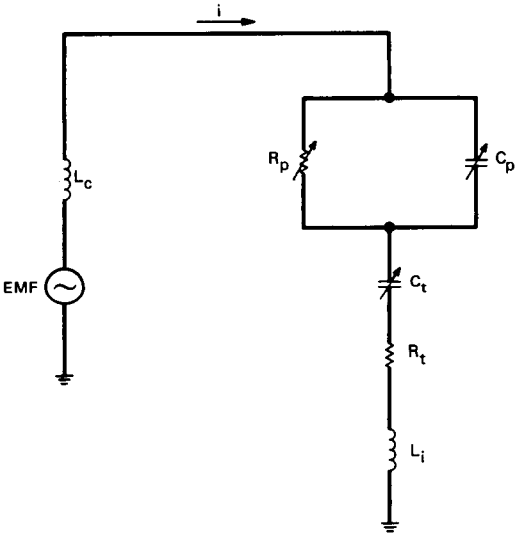


Fig. 10. Q modulator for varying C_p

Traveling-Wave Maser Closed-Cycle Refrigerator Data Acquisition and Display System

L. Fowler and M. Britcliffe
Radio Frequency and Microwave Subsystems Section

This report describes a data acquisition and display system that automatically monitors the performance of the 4.5-K closed-cycle refrigerators used to cryogenically cool traveling-wave masers. The system displays and stores operating parameters for the purpose of providing status information, failure prediction, and analysis. A prototype of this system will be installed at DSS 12 in the near future. The advantages of using commercial data acquisition hardware with installed operating systems and BASIC programs for this application are discussed.

I. Introduction

Traveling-wave maser (TWM) closed-cycle refrigerator (CCR) systems have been used in the DSN for over 20 years. The CCR system consists primarily of a helium compressor and a refrigerator assembly (Fig. 1). The purpose of the system is to maintain the temperature of the low-noise maser amplifier assembly at 4.5 K. Maintaining this temperature is critical to the proper operation of the maser.

Closed-cycle refrigerator failures are responsible for a large percentage of unscheduled maintenance time in DSN tracking operations. The mean time between failures is relatively short (2500–3000 hours), and the time to return to operation after a failure is long (up to a few days). It would be very advantageous to improve the operator's ability to observe CCR conditions so that failures could be better diagnosed and so that imminent CCR failures could be predicted and maintenance scheduled in advance. Some CCR failures can be predicted by careful analysis of CCR performance data.

The monitoring of CCR operation is currently performed manually by station personnel on a daily basis. This involves reading a number of gauges and transducer readouts for each TWM/CCR system and recording the information in a logbook. These instruments are located at four separate locations at the antenna site (Fig. 2).[1] This data must then be analyzed by engineering personnel at the station. Unfortunately, this time-consuming and tedious task rarely results in timely failure prediction or accurate failure diagnosis.

Another problem associated with CCR monitoring is a lack of system operating-temperature data. Most DSN instal-

[1] *Operation and Maintenance, Traveling-Wave Maser Group, Block III, 26-Meter Antenna*, JPL Publication TM 00714 (internal document), Jet Propulsion Laboratory, Pasadena, California, May 3, 1976; and *X-Band TWM Equipment, Block I (34-Meter Antenna)*, JPL Publication TM 511967 (internal document), Jet Propulsion Laboratory, Pasadena, California, February 2, 1979.

lations have no provision for monitoring the operating temperature of the compressor or displaying temperature data from cryogenic temperature transducers installed in the refrigerators.

The TWM/CCR data acquisition and display system will provide a performance data base of both normal and abnormal operating characteristics, enabling engineering personnel to evaluate CCR failure dynamics and to develop failure prediction algorithms. The operator interface will be one central terminal (personal computer and printer), located in the maintenance facility. Maintenance personnel will be automatically notified of out-of-limit conditions, and performance data will be logged automatically.

II. Description

The CCR data acquisition and display system consists of refrigerator and compressor sensors and data acquisition assemblies, a communications and display processor, and communications expanders. CCR monitor equipment locations are shown in Fig. 2, and a block diagram of the system configured for DSS 12 is shown in Fig. 3. The DSS-12 system monitors three compressors and three refrigerators; each system is capable of monitoring four refrigerators and four compressors.

A. Sensors

Performance sensors in the present system are quite limited. They consist of gauges and transducers that must be read by an operator. Pressure data is obtained from conventional gauges located on the compressor. No provision is made for measurement of the compressor operating temperatures or compressor motor current.

JT flow is the only indicator of refrigerator performance that can be measured remotely from the refrigerator package. It is measured with a Hastings flowmeter installed near the compressor. Cryogenic temperature readings are limited in most cases to a vapor pressure thermometer mounted on the refrigerator package that indicates the vapor pressure of liquid hydrogen in a bulb mounted on the 4.5-K stage.

The system functions to be monitored in the new system were selected primarily to provide the information necessary to verify proper operation and predict failures. Most of the existing manual sensors are monitored in parallel. Additional sensors were chosen to monitor specific trouble points and to aid in the identification of long-term failure modes.

Another sensor selection and design criterion was the ability to install the sensors on a helium compressor in the field without opening the helium piping or altering the existing wiring, hardware, etc. The complete compressor sensor package can be installed on an operational compressor in a matter of minutes without warming the refrigerator. A comparison of sensors used on the current system and the new automated system is given in Table 1.

Helium pressure data is obtained from strain gauge-type pressure transducers connected to the CCR using self-sealing quick-disconnect fittings. Compressor temperatures are measured using Analog Devices monolithic temperature sensors connected to the outside of the helium lines. Motor current is measured using current transformer-type sensors connected in-line with the compressor power cable. The JT flow data is obtained from an analog output connection on the existing system flowmeter. The cryogenic temperature measurements are made using Lake Shore Cryotronics silicon diode thermometers that are being installed on DSN TWM/CCR systems as opportunity permits.

B. Passive Refrigerator Capacity Monitor

The system uses a real-time refrigeration capacity monitor based on a concept developed at JPL and implemented on early R&D CCRs [1]. Unlike most methods of measuring capacity, no heat is applied to the refrigerator, and therefore it can operate continuously. An operator can verify the relative health of the refrigerator in one glance. The new monitor has demonstrated an accuracy of 5 percent in laboratory tests.

C. Refrigerator and Compressor Data Acquisition Assemblies

The data acquisition assemblies for the refrigerators and compressors consist of a commercial Analog Devices data acquisition module, interface hardware, and a power supply. The hardware is the same for both the refrigerator and compressor data acquisition assemblies, with the exception that a current source (PC board) is added to the refrigerator assemblies to excite the cryogenic temperature transducers. The Analog Devices μMAC-5000 module is an 8088-based microprocessor with a 12-channel analog-to-digital converter and signal conditioning circuitry. The module provides 46 kbytes (later versions provide 56 kbytes) of user memory, two serial communications ports, and a socket for a user-installed EPROM (to contain the user's program). The software (resident in the EPROM) is the same for both the refrigerators and the compressors—the program is customized for either application by setting a DIP switch.

One data acquisition processor is used for each compressor, since all 12 available channels are required. The refrigerators

require only five channels each, so one data acquisition assembly is used to monitor two refrigerators.

D. Communications and Display Processor

All communications and displays are controlled by the host processor, which is an IBM PC-XT microcomputer. The IBM standard motherboard has been replaced by a CTXT Terminator 286 board, and the standard 60-W power supply has been replaced by a 130-W power supply. In addition, a 30-megabyte hard disk drive and a hardware clock with an additional serial port have been installed. The hard disk holds both the accumulated data and the operating programs. The floppy drive is used to transfer both data and programs to and from one personal computer to another. As new versions of the program are developed, they can be easily installed in existing systems. Data can also be transferred from the TWM/CCR personal computer to other personal computers for further processing by engineering personnel, as is planned for the Parkes TWM/CCR implementation.

The displays and operating software will be described in detail in a separate report. Basically, each measured parameter of the system is displayed continuously (see Fig. 4). An out-of-limit condition is indicated with a reverse background screen around the data. A communications failure is also indicated by a reverse background screen around the applicable equipment heading. The operator can create a printout of status information and reset the limits using simple commands. A complete set of data is written to the hard disk automatically every 15 minutes. The hard disk is capable of storing up to 1 year of data.

E. Communications Expanders

Prototype communications expanders (JPL PN 9489463) were used to hook together four RS-232 ports into a wire-OR configuration (see Fig. 5). Each expander also establishes a current loop configuration to link the previous expander (or data acquisition assembly) to the next expander (or data acquisition assembly). The communications expanders were used to provide reliable long-distance communications (the RS-232C communication ports provided by the personal computer and the data acquisition modules specify 25 feet maximum). In addition, the expanders convert single-channel communications to multi-channel (up to four channels per expander). Standard eight-pin connectors are used for all ports.

III. Plans

The present data acquisition system represents Phase I of the automation plan. Phase II will expand monitor capabilities to enable remote fetching of data and a program for plotting the data. In addition, CCR system monitoring could be expanded to include maser electronics (pump sources and the superconducting magnet). The basic system could also be adapted to include control of the pump sources and magnet and of the automated cool-down control system. The system can also be modified to monitor HEMT or FET amplifier refrigerators as they are implemented into the DSN.

IV. Conclusion and Recommendations

The Phase I data acquisition system has gone through a long evolutionary cycle. The original design concept was to use CCM standard multibus modules (DSN standard practice) for both the communications and display processor and the data acquisition processors. After funding for this effort was reduced, it was decided to eliminate the control function from the equipment for the initial phase and to investigate commercial off-the-shelf equipment which had recently become available.

It was found that the Analog Devices data acquisition module had all the features required for the TWM/CCR application. An added benefit was the high-quality BASIC programming language standard on the module. This program is interactive rather than compiled (as is the case for PLM) and resulted in increased ease of programming. Additionally, the size of the packaged module is about one-half the size of the CCM module package, an important factor in the already crowded compressor and refrigerator areas. Also, commercial manuals are supplied, substantially reducing the effort required to produce a DSN operations and maintenance manual.

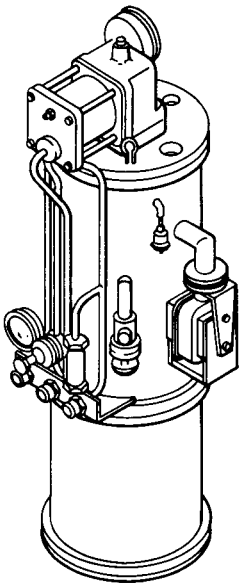
The benefits resulting from selection of the personal computer communications and display processor are similar to the data acquisition assemblies. The equipment is low cost, development time was cut drastically because the operating system was already built in, and the effort required to produce manuals is significantly reduced.

Reference

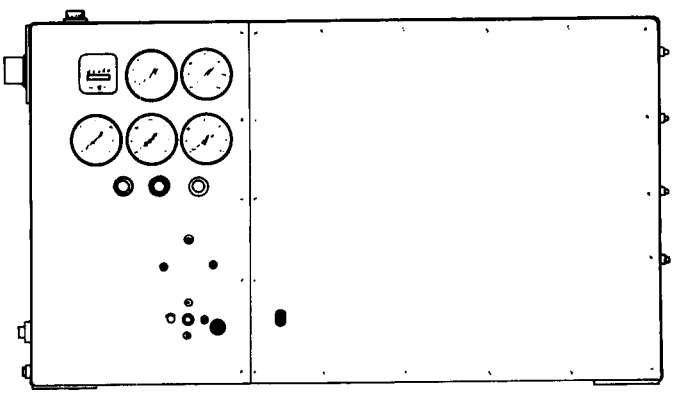
[1] W. H. Higa and E. Weibe, "One Million Hours at 4.5 Kelvins," *National Bureau of Standards Technical Publication No. 508*, pp. 99–107, April 1978.

Table I. Comparison of sensors on old system versus new system

Sensor function	Sensor type	
	Old system	New system
Compressor supply pressure	Gauge	Transducer
Return pressure from 1st and 2nd stage of refrigerator	Gauge	Transducer
Return pressure from JT stage of refrigerator	Gauge	Transducer
Compressor storage tank pressure	Gauge	Transducer
Pressure drop in compressor oil separator	None	Transducer
Temperature of compressor 1st stage	None	Thermometer
Temperature of compressor 2nd stage	None	Thermometer
Temperature of compressor motor	None	Thermometer
Compressor motor AC current	None	Transducer
JT circuit helium flow rate	Flowmeter	Flowmeter
Temperature of refrigerator 4.5-K stage	VP gauge	Thermometer
Temperature of refrigerator 2nd stage	None	Thermometer
Temperature of refrigerator 1st stage	None	Thermometer
Refrigerator cooling capacity	None	Monitor



CLOSED-CYCLE REFRIGERATOR



HELIUM COMPRESSOR

Fig. 1. Basic components of the closed-cycle refrigerator system

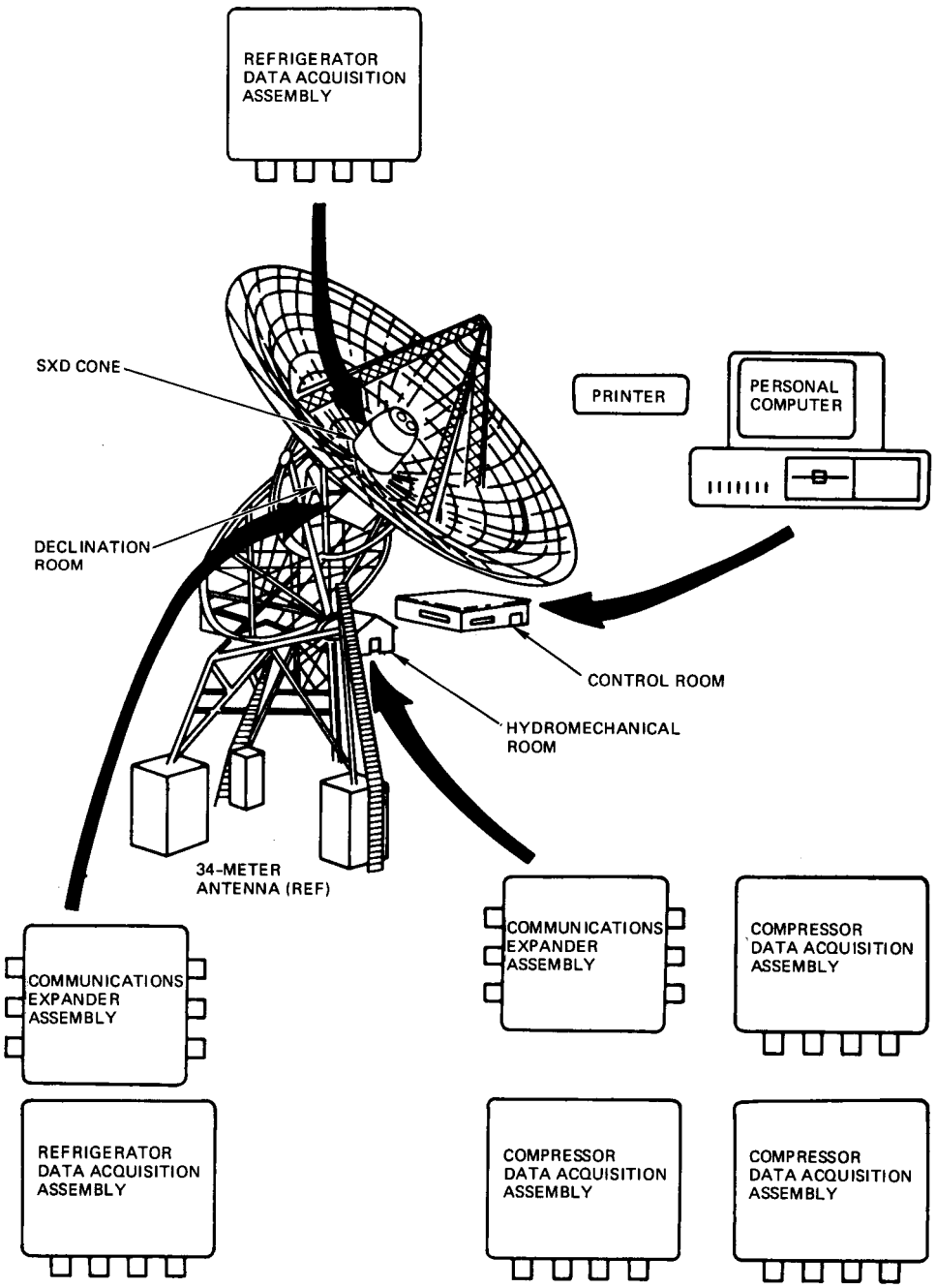


Fig. 2. CCR data acquisition equipment locations

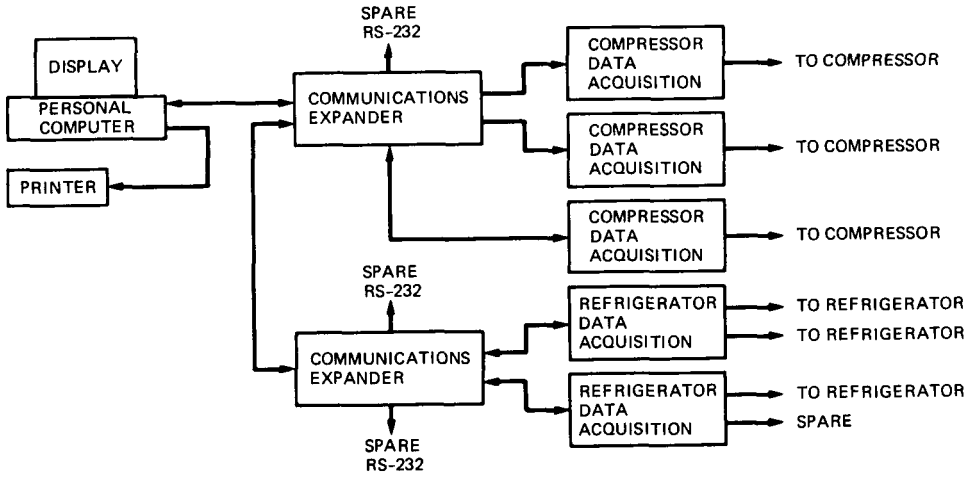


Fig. 3. CCR data acquisition system block diagram

Parameter #1 #2 #3 #4 Units

System Communications	R	C	R	C	R	C	R	C	Status
4.5° Stage Temperature	4.62		4.97		-*-		-*-		Deg K
Heat Exchanger Temp.	12.7		13.8		-*-		-*-		Deg K
15° Stage Temperature	15.8		16.0		-*-		-*-		Deg K
70° Stage Temperature	68.1		67.5		-*-		-*-		Deg K
Reserve capacity	55.0		20.6		-*-		-*-		%
Vacuum	7E-8		9E-7		-*-		-*-		Torr
JT Flow	1.47		1.38		-*-		-*-		Scfm
Supply Pressure	245.		234.		-*-		-*-		Psi
Refrigerator Return	97.6		100.		-*-		-*-		Psi
JT Return Pressure	3.26		5.09		-*-		-*-		Psi
Storage Tank Pressure	185.		146.		-*-		-*-		Psi
Oil Separator Delta	15.8		14.3		-*-		-*-		Psi
Motor Temperature	83.4		87.7		-*-		-*-		Deg C
1st Stage Temperature	85.9		88.2		-*-		-*-		Deg C
2nd Stage Temperature	86.7		88.3		-*-		-*-		Deg C
AC Input Current φA	11.6		10.5		-*-		-*-		Amps
AC Input Current φB	13.3		12.0		-*-		-*-		Amps
AC Input Current φC	12.8		12.1		-*-		-*-		Amps

232:14:54:04 Thursday, August 20, 1987 Press F10 for Menu.

Fig. 4. Sample display

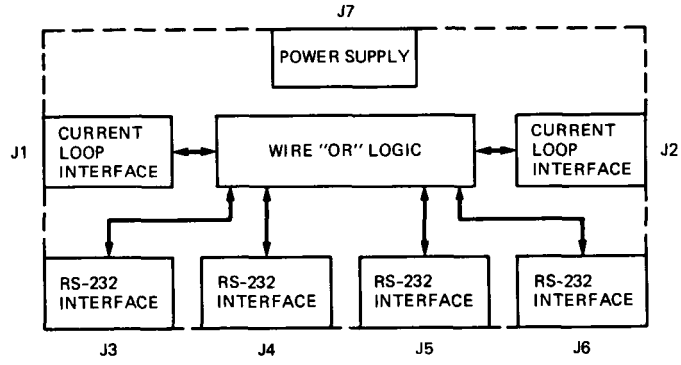


Fig. 5. Communications expanded block diagram

Two-Watt, 4-Kelvin Closed Cycle Refrigerator Performance

M. Britcliffe
Radio Frequency and Microwave Subsystems Section

This article describes a 2-watt, 4-K helium refrigerator using the Gifford–McMahon/Joule–Thomson cycle. The unit features a removable displacer cylinder and high-efficiency, low-pressure-drop heat exchangers. These improvements result in a 100 percent increase in cooling power over the existing DSN system. The effects of heat exchanger efficiency and Gifford–McMahon expander performance on refrigerator cooling capacity are also discussed.

I. Introduction

The 1-watt, 4-K refrigerator currently used to cool DSN traveling-wave masers was developed at JPL over 20 years ago [1]. It represented an order-of-magnitude improvement in reliability and performance over the commercial unit used previously. These systems have been used with only minor modification since that time.

Over the years, several problems have become apparent with the existing DSN design. Current DSN masers have grown to require more cooling power. The original 2.3-GHz maser required only 150 milliwatts of the 1 watt available, leaving a reserve of 850 milliwatts. Some current designs reduce the reserve to 500 milliwatts. Experience in the DSN has shown that systems with reduced capacities have shorter mean times between failures.

Many DSN refrigerators have been in operation long enough to experience a reduction in cooling capacity due to wear of the Gifford–McMahon (GM) expansion engine displacer cylinder. Replacement of the cylinder requires moving the entire Joule–Thomson (JT) circuit, which is a costly and time-consuming operation.

The current DSN refrigerator is also difficult and expensive to produce. Fabrication is very labor intensive, and some materials are difficult to obtain.

The cost of the refrigerator alone is a very significant part of a complete maser system. The new 2-watt refrigerator described in this article features substantially increased cooling capacity, simpler heat exchangers with improved efficiency, and a conveniently replaceable GM displacer cylinder. The unit will accept all current maser assemblies without modification. The physical size of the refrigerator and the input power for the system remain unchanged.

II. Thermodynamic Considerations

Both the existing DSN 1-watt system and the new 2-watt system described here use a combination of the Gifford–

McMahon (GM) and Joule-Thomson (JT) thermodynamic cycles. A gas flow schematic is shown in Fig. 1.

A. Cycle Description

The refrigerator uses a two-stage CTI Model 350 GM expansion engine that simultaneously provides 25 watts of cooling at 60 K and 5 watts at 15 K to precool helium gas flowing in the JT circuit. The JT circuit consists primarily of three heat exchangers and a JT expansion valve. High-pressure helium (300 psig) supplied by a compressor at ambient temperature (300 K) is cooled in the heat exchangers by the gas returning from the cold station and by the expansion engine. When the gas reaches the JT expansion valve, it has been cooled to nearly the temperature of the 4-K stage. The JT valve is simply a calibrated restriction that allows the high-pressure gas to expand into the low-pressure JT return line. During this expansion, the helium cools slightly, and a fraction of the flow condenses into 4-K liquid. Heat from the load—in this case, the maser—is absorbed by vaporizing this liquid.

B. Heat Exchanger Efficiency

Heat exchanger efficiency is one of the most critical aspects in the design of a JT refrigerator. It determines the amount of heat that must be removed from the supply helium by the GM engine. It also determines the helium flow rate required in the JT circuit to produce a given amount of cooling at 4 K. Heat exchanger efficiency is defined by

$$e = \frac{q_{\text{actual}}}{q_{\text{max}}}$$

where q_{max}, the maximum amount of heat that can be transferred to or from the helium, is a function of the fluid properties of the helium for a given temperature and pressure. The amount of heat actually transferred, q_{actual}, is primarily a function of the design of the heat exchanger (heat transfer area, material thermal conductivity, flow passage size, etc.).

The significance of heat exchanger efficiency can be seen by the fact that the production of 1 watt of cooling at 4K requires approximately 130 watts of heat to be transferred from the incoming gas upstream of the JT valve. Any heat not removed by the heat exchangers must be absorbed by the engine. Because the amount of heat to be removed from the helium is large compared to the capacity of the engine, the numeric efficiency must be high. Table 1 shows the estimated heat exchanger efficiencies for both systems.

Pressure drop in the return path of the heat exchangers is also important. The operating temperature of the 4-K stage is determined by the helium pressure at the cold station. Any pressure drop in the JT return line will increase the pressure, and therefore the temperature, of the 4-K stage.

C. Gifford–McMahon Expansion Engine Performance

Expander performance is another crucial factor in GM/JT refrigerator operation. As stated earlier, the expansion engine absorbs any heat not removed from the supply gas stream before it reaches the final-stage heat exchanger. The engine also cools the radiation shield that intercepts thermal radiation from ambient-temperature sources.

The expander is a reciprocating mechanical device with several moving parts that are subject to wear. As this wear occurs, the efficiency of the engine decreases and the operating temperatures of the engine stages increase. This increase in engine operating temperatures has a dramatic effect on the 4-K cooling capacity of the refrigerator.

Another factor that affects engine performance is external heat load from sources other than the helium gas in the JT circuit. Thermal radiation from room temperature is intercepted by radiation shields cooled by the engine first stage. Radiation load from large shields can equal the load from the helium. Heat conduction through waveguides, supports, and wiring also contributes to engine load.

III. Hardware Description

Both the DSN 1-watt refrigerator and the new 2-watt refrigerator are shown in Fig. 2. They are nearly identical in physical size and weight. The new refrigerator was designed to accept existing maser hardware without modification. Many of the components developed for the initial design are used in the new system. The most significant changes are the use of improved efficiency heat exchangers and the addition of a "bolt-in" expander cylinder.

The new heat exchangers are based on a concept developed at JPL [1] and later refined at the National Radio Astronomy Observatory. They consist of a spiral coil of convoluted tubing wound on a Micarta mandrel and enclosed in a thin stainless steel tube (Fig. 3). The supply helium travels through the inside of the convoluted tubing, and the return gas passes axially over the outside. This new design contains more than three times the heat transfer area of the original design.

The expander–displacer cylinder is a commercial unit supplied by CTI. The heat stations are bolted to flanges on the cylinder rather than being soldered. The cylinder is bolted to the vacuum housing end plate using an O-ring seal. The cylinder is shown removed from the JT circuit in Fig. 4.

Fabrication costs and assembly time have been reduced by as much as 50 percent on the new refrigerator. Fabrication of the original displacer cylinder involved several machining, welding, heat treatment, and inspection processes that resulted in high cost and long delivery times. Interestingly, the entire 2-watt refrigerator was built and tested while waiting for two of the original-design cylinder assemblies to be fabricated. Eight solder or weld joints have been eliminated from each heat exchanger. Joints of this type are often the cause of internal helium leaks that plague the construction of helium refrigerators. The materials used in the construction of the heat exchangers are easily obtainable.

IV. Performance

Performance specifications for the existing system and the new design are shown in Table 2. The 4-K cooling capacity of the new unit is more than double that of the existing design. The input power required remains unchanged. This results in a 100 percent increase in thermodynamic efficiency for the total closed-cycle refrigerator system.

Figures 5 and 6 represent the effects of first- and second-stage engine loading on 4-K capacity. The data shows a substantial improvement in the refrigerator's resistance to external heat loads.

V. Conclusions and Recommendations

Cooling capacity at all three stages of refrigeration has been improved substantially in the new 2-watt system. This results in an increase in the refrigerator's resistance to GM expander performance degradation and external heat loads. The extra cooling power should result in a marked improvement in DSN traveling-wave maser closed-cycle refrigerator-system mean time between failures. The reduction in fabrication expenses makes the unit a cost-effective replacement for aging 1-watt systems.

Future projects such as the 32-GHz maser and masers with cryogenically cooled feeds may require the extra cooling this refrigerator provides. HEMT amplifiers can also be cooled to 4 K to reduce system noise temperature. Although the HEMT device itself does not benefit greatly from 4-K operation, thermal noise contribution from microwave components at the device's input (e.g., from filters and isolators) can be reduced by lowering their physical temperature.

Acknowledgment

The author would like to thank T. Hanson of Bendix Field Engineering for providing the superb craftsmanship that went into the fabrication of the prototype refrigerator. He was also responsible for the initial cryogenic testing of the unit.

Reference

[1] W. H. Higa and E. Wiebe, "One Million Hours at 4.5 Kelvins," *National Bureau of Standards Technical Publication No. 508*, pp. 99–107, April 1978.

Table 1. Estimated heat exchanger efficiency

Heat exchanger	1-watt CCR	2-watt CCR
First stage	0.94	0.96
Second stage	0.93	0.96
Third stage	0.97	0.99

Table 2. Comparison of heat exchanger specifications

Specification	1-watt CCR	2-watt CCR
Operating temperature	4.5 K	4.4 K
Cooling capacity at operating temperature	0.95 W	2.2 W
JT mass flow	1.5 SCFM	2.3 SCFM
Compressor input power required	8000 W	8000 W
Net thermodynamic efficiency	8400 W/W	3600 W/W

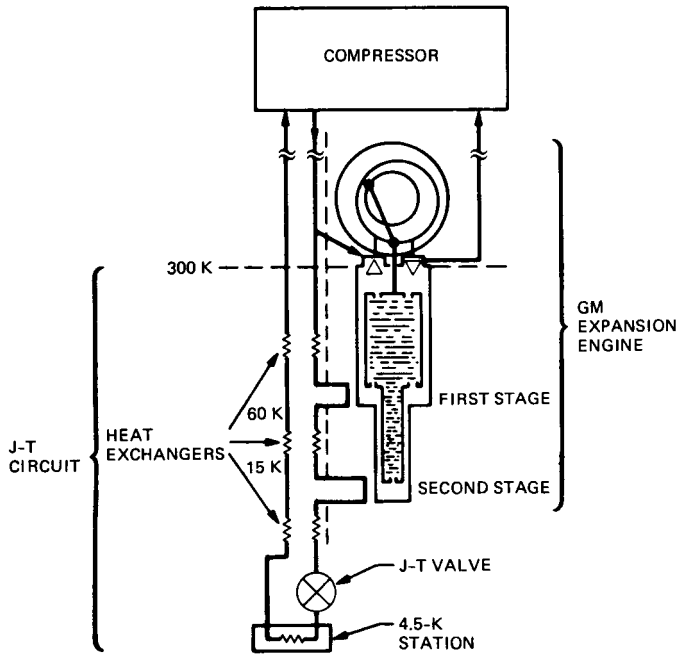


Fig. 1. GM/JT refrigerator gas flow schematic

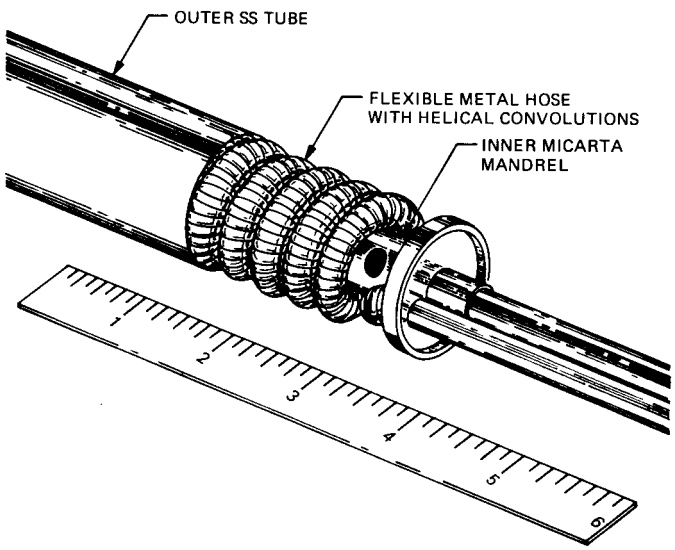


Fig. 3. New heat exchanger

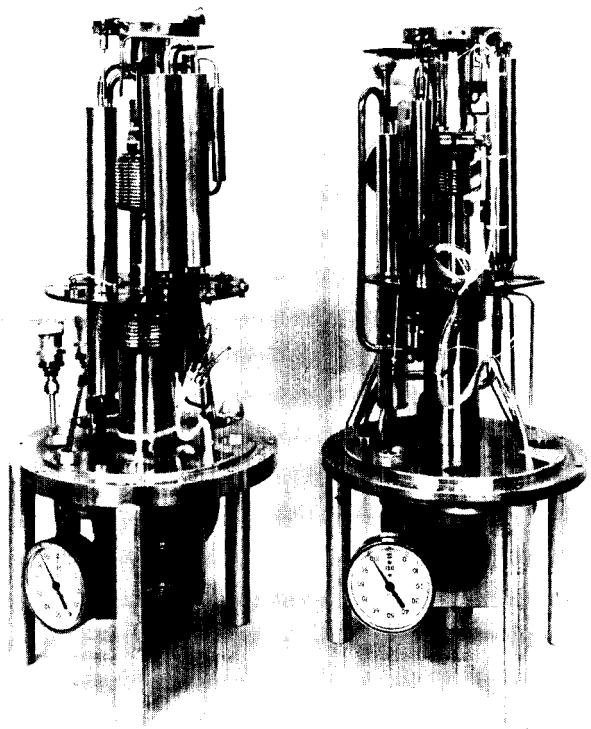


Fig. 2. Two-watt refrigerator (shown on right) and 1-watt refrigerator

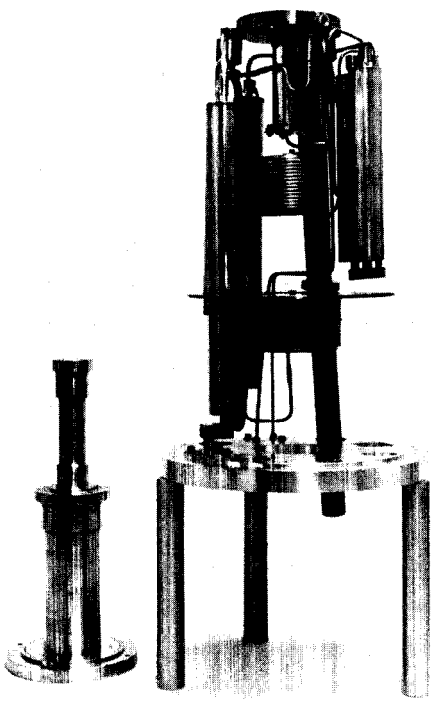


Fig. 4. Displacer cylinder removed from refrigerator, leaving JT circuit intact

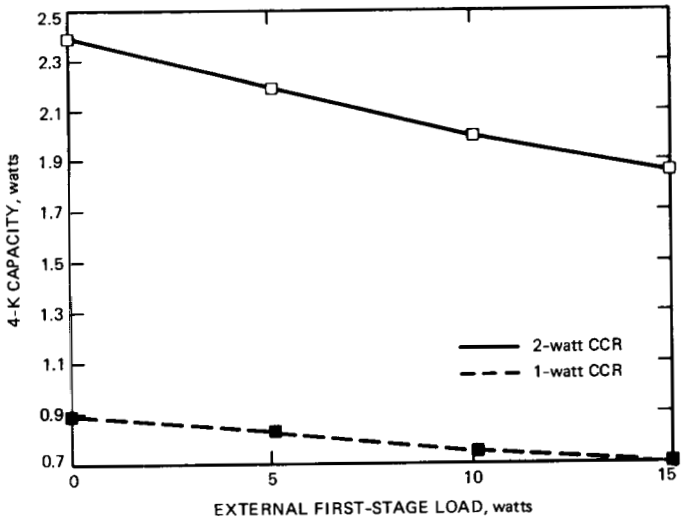


Fig. 5. Four-kelvin capacity versus first-stage load

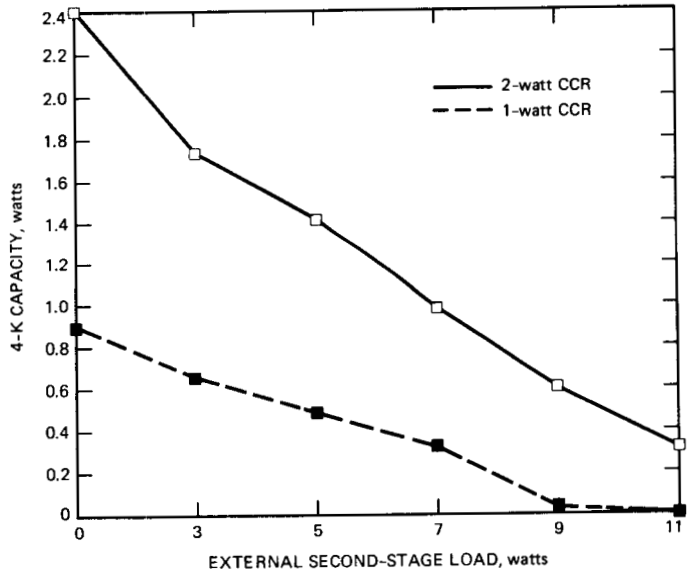


Fig. 6. Four-kelvin capacity versus second-stage load

Tau Ranging Revisited

R. C. Tausworthe
Information Systems Division

The advent of reasonably fast and inexpensive Very Large Scale Integration (VLSI) components offers an opportunity to reconsider the merits of composite-code-uplink ranging systems, abandoned years ago. It is shown in this article that a ranging receiver with a sufficient and reasonable number of correlators is competitive with the current sequential component ranging system and may outperform that system by some 1.5–2.5 dB. The optimum transmitter code, the optimum receiver, and a near-maximum-likelihood range-estimation algorithm are presented.

I. Introduction and Background

Prior to 1973, planetary and lunar spacecraft ranging systems at JPL utilized a transmitted uplink code made by combining binary clock and pseudonoise sequences in a majority-vote logic [1].[1] The "composite-code-uplink" ranging receiver consisted of one or two channels that correlated the transponded signal with combinations of the clock and each separate component sequentially through each successive symbol-delay to determine the precise delay on each of the components. Since transmitted power was distributed among the various clock and pseudonoise code components but the receiver was sensitive to only one component at one phase at a time, acquisition time was longer by about a factor of 16 than if the receiver could have processed all the received power during the acquisition time.

Consequently, once sufficient analysis and precautions had been taken to ensure that uplink ranging sidebands would not interfere with the spacecraft command system, a "sequential-component-uplink" ranging method [2] was devised that is still being used today. The term "sequential," in this case, refers to the transmitted code, which is a time series of square waves of successively shorter wavelengths. The receiver, still a few-correlator device, is programmed to acquire the sequentially transmitted components one by one. The newer method had a 16:1 acquisition-time advantage over the older scheme because it could utilize all the transponded power for each component during the acquisition time. The necessity to program the uplink code and to program the receiver to switch components at the proper round-trip-light-time interval was a disadvantage compensated by the signal-to-noise advantage.

During a transition period, both ranging techniques were used. The older, composite transmitted code method was referred to as the "τ" system, and the newer, sequential component code method was called the "μ" system. These names were dubbed by Robertson Stevens, now the Chief Engineer

[1] R. W. Tappan and R. C. Tausworthe, "DSIF Technical Description, Planetary Ranging Equipment, Mariner-Mars 1971 Configuration," JPL Document TD505943A (internal document), Jet Propulsion Laboratory, Pasadena, California, February 4, 1972.

of the Deep Space Network, one day while strategizing at the blackboard in a design meeting, groping for a notation to distinguish the two. They derive from the initials of the then-purveyors of the two systems: the author (guess which one) and Warren Martin. Usage of these designations has decayed over the years because the τ system is no longer extant. The designations are reinstituted in this article for brevity in referencing the two schemes.

The τ planetary composite transmitter code was generated by combining a clock square wave with a majority-vote logic of 5 pseudonoise sequences in an exclusive-OR fashion. The components had symbol-periods of 2, 7, 11, 15, 19, and 23, for a total code period of N = 1,009,470. This, when clocked at a symbol-period t_0 of about 1 μsec, gave a repetition period of about 1 sec, yielding a 2-way range ambiguity interval of approximately 150,000 km.

The majority-vote combining logic in the old-τ system was chosen [3], [4] because it evenly (and optimally, for that strategy) distributed the power among all components for sequential detection. All component delay measurements were thus made with times approximately proportional to the component periods. The clock-code (period 2) phase was acquired first, and then the other 75 code phase correlations were made sequentially. Two correlator channels were time-shared during range acquisition in a way that balanced the channel gains and removed any residual dc bias voltage in the baseband detection process.

The reason for utilizing only a few correlators in ranging receivers until now has been that each correlator channel consisted of relatively expensive analog and unit-logic digital hardware. However, with the advent of reasonably high speed digital analog-to-digital devices and very large scale integrated (VLSI) digital devices, correlators may now be made at much more modest cost [7].

It is now economically feasible under VLSI technology to put the needed number of correlators into a receiver to build a detector for each of the components at each symbol-delay of the composite code. (It still may be impractical to build a full matched filter for the overall transmitted code, however.) It is therefore appropriate to reevaluate the relative merits of the two ranging methods under the removed constraint that previously fixed the number of correlator channels.

This article shows that a new-τ composite-component uplink code (or "$\nu\tau$," for brevity) which utilizes a new combining logic for the transmitter code and a 77-correlator receiver is again favorable in performance. In fact, it is shown that the $\nu\tau$ method is only about 0.25 dB below the performance of a matched filter for the optimal transmitter code. As the μ system is now configured, about half of the range-measurement time is spent in correlating with the highest-frequency component (the clock), the other half being spent determining the range-cells of the lower-frequency components. The $\nu\tau$ system thus outperforms it by some 2.5 dB in signal-to-noise ratio.

A companion paper by the author and J. R. Smith [7] discusses the requirements and conceptual design of the correlator channels and VLSI devices required.

II. The Transmitted Code

The code component periods remain the same as in the earlier τ system: 2, 7, 11, 15, 19, and 23. The combining logic for the transmitted code, however, is now taken to be

$$x(\mathbf{c}) = c_1 \text{ XOR AND } (\mathbf{c}) \tag{1}$$

in which $\mathbf{c} = (c_1, \ldots, c_6)$, c_1 is the clock (period 2) sequence, XOR is the exclusive-OR function, and AND ( ) is the logical-AND of all 6 component sequences. The logical-AND, or "unanimous-vote" logic, is the limiting case of the majority-vote logic used previously.

Since AND (c) contains only one "1" in its truth table of 64 entries, the in-phase cross-correlation of $x(\mathbf{c})$ with c_1 will be [3]

$$R_{x1} = \frac{(64 - 1) - 1}{64} = 0.97 \tag{2}$$

The in-phase cross-correlation of $x(\mathbf{c})$ with a code of the form $(c_1 \text{ XOR } c_i)$, for $i = 2, \ldots, 6$, is only

$$R_{xi} = \frac{1 + 1}{64} = 0.031 \tag{3}$$

(These figures are only approximate, being influenced slightly by the sense of imbalances between 0's and 1's in each of the pseudonoise sequences. These imbalances can be chosen to further optimize the reception, but this is left as an exercise for the implementer.)

The cross-correlations as functions of ranging delay are shown in Fig. 1.

III. Performance

The $\nu\tau$ receiver achieves range measurement precision by clock-component correlation, just as did both predecessors. But since the clock component of the transmitted code contains 94 percent (i.e., 0.97^2) of the total ranging power, there

is only a 0.27 dB degradation in acquisition time from transmitting the clock component alone (such a code would not, however, remove the ambiguity of the range).

The requirements for range accuracy demand that the standard deviation of the range measurement due to noise be about 1/1000 of the "chip" time, t_0, or symbol-rate of the clock. The relative clock variance σ_Δ^2 of $\Delta = \tau/t_0$ is thus about 10^{-6}. In order to achieve this accuracy, the received signal-energy/noise-density ratio must accordingly be high [5].

$$\frac{ST}{N_0} = \frac{1}{16 R_{xi}^2 \sigma_\Delta^2} = 6.7 \times 10^4 \tag{4}$$

where S is the total signal power, T is the correlator integration time, and N_0 is the received noise (single-sided) spectral density.

Detection of the pseudonoise component phase is accomplished by correlating the received signal with each of the separate phases of (c_1 XOR c_i). The power in each component is only about $0.001 = 0.31^2$ of the total, so the component detection-energy/noise-density ratio is about

$$\frac{S_i T}{N_0} = \frac{R_{xi}^2 ST}{N_0} = 0.001 \times 6.7 \times 10^4 = 67 \tag{5}$$

The maximum required pseudonoise component-energy/noise-density ratio for an error probability of 0.01 was about 10 for the old-τ system. Since detection of the pseudonoise range cell (see below) involves summing channel values for *pairs* of correlators, the additional noise may degrade the required $S_i T/N_0$ to about 20 (a full analysis has not yet been made). The better-than-a-factor-of-three margin ensures, however, that unerring full-range acquisition is almost certain.

IV. Maximum-Likelihood Receiver

The maximum-likelihood estimator of the clock component phase is well known and will not be repeated here. The remaining pseudonoise code delays, however, are to be estimated from measurements made in parallel with the clock delay determination. This approach differs significantly from the old-τ method, where the receiver pseudonoise sequences were acquired *after* the clock so that the receiver codes could be adjusted to then be in step with the received signal. Thus, whereas the old-τ system enjoyed full component correlation in only one integration bin per component, the $\nu\tau$ scheme must make do with partial component correlation in two adjacent correlation channels for each component.

The derivation of the maximum-likelihood detector is straightforward: We presume that we receive the transmitted binary ranging signal $x(t) = x(\mathbf{c}[t])$, normalized here to unit power, immersed in wideband Gaussian noise $n(t)$, as

$$y(t) = \alpha x(t - \tau) + n(t) \tag{6}$$

where $\alpha = S^{1/2}$. The time delay τ is to be estimated as that value $\hat{\tau}$ maximizing the conditional probability (density) function

$$p\{\tau \mid y(t), 0 \leqslant t \leqslant T\} \tag{7}$$

Under the usual assumption that τ is uniformly distributed over the unambiguous-range interval, the likelihood ratio, by Bayes' rule, is

$$\lambda = \frac{p\{\hat{\tau} \mid y(t)\}}{p\{\tau \mid y(t)\}} = \frac{p\{y(t) \mid \hat{\tau}\}}{p\{y(t) \mid \tau\}} \geqslant 1 \tag{8}$$

where the interval $(0, T)$ dependency has been suppressed for notational convenience.

The probability of receiving $y(t)$, given τ, is the probability that the noise in $n(t) = y(t) - \alpha x(t - \tau)$. Because of the wideband Gaussian character of the noise, the likelihood ratio becomes [6]

$$\lambda = \frac{\exp\left(-\frac{1}{N_0}\int_0^T \left[y(t) - \alpha x(t - \hat{\tau})\right]^2 dt\right)}{\exp\left(-\frac{1}{N_0}\int_0^T \left[y(t) - \alpha x(t - \tau)\right]^2 dt\right)} \geqslant 1 \tag{9}$$

where $\exp(x)$ is the exponential function e^x.

By noting $x^2(t) = 1$, canceling like terms in the numerator and denominator, and taking logarithms, we find that the condition on $\hat{\tau}$ is that

$$\int_0^T y(t) x(t - \hat{\tau})\, dt \geqslant \int_0^T y(t) x(t - \tau)\, dt \tag{10}$$

That is, $\hat{\tau}$ will be the maximum-likelihood estimator of τ provided that it maximizes the correlation between the observed $y(t)$ and the delayed transmitted code. However, a continuum of correlators is infeasible, so we must infer $\hat{\tau}$ from the finite number of measurements we do make. It has been shown [3] that the maximum-likelihood value can be inferred using correlations of the incoming signal and various delays of

the transmitted codes, in the form $x_1 = c_1$ and $x_i = c_1$ XOR c_i for $i = 2, \ldots, 6$.

$$I_{ij} = \int_0^T y(t) x_i(t - \hat{\tau}_j)\, dt \tag{11}$$

To decrease ranging inaccuracy caused by waveform distortion within the communication system, estimation of the clock component phase in the current μ system is performed by maximum-likelihood methods applied only to the fundamental harmonic of the received clock component. This results in a modest, justifiable increase in required integration time. The $\nu\tau$ method would presumably have the same requirement for this clock estimation scheme.

Since almost all of the transmitted power is in the clock component of the code, the contribution of the 75 other correlators in improving the accuracy of the clock phase estimate will be negligible. Hence, the value of $(\hat{\tau} \bmod 2t_0)$ may be determined from the clock-channel correlators alone. (Combined, weighted estimation of the clock phase from all channels can be done, however, with only slightly more complexity in the estimation program, if desired.) Since 94 percent of the transmitted power is in the clock component, and since maximum-likelihood estimation is performed on this component, the clock phase estimate is very nearly the same as the maximum-likelihood estimate of a pure clock signal.

Thus, any method that with high likelihood selects the proper range-cell delays of the remaining components will measure the range within 0.27 dB of the performance of a maximum-likelihood device, insofar as ranging accuracy is concerned.

The clock-channel measurement of $(\hat{\tau} \bmod 2t_0)$ is required to be very close to the actual value of $(\tau \bmod 2t_0)$ for system accuracy. From this value, $(\hat{\tau} \bmod t_0)$ may be determined, as well as the ± 1 sense of the in-step correlation (Fig. 1). Therefore, determination of $(\hat{\tau} \bmod Nt_0)$ additionally requires only the estimation of the integer values k_i such that

$$\hat{\tau} - k_i t_0 = (\hat{\tau} \bmod t_0) \tag{12}$$

for each of the remaining pseudonoise components (period $N_i t_0$). We may estimate these values from the correlator outputs of each pseudonoise component and combine them to form the overall range using the Chinese remainder theorem, just as did the previous τ system.

Only two of the correlator integration values I_{ij}, $j = 1, \ldots, N_i$ for the ith code component may derive from partial correlation with the true delay. The other I_{ij} values correspond to out-of-phase correlation levels. Let z_{ij} denote the normalized sum of adjacent correlators at the jth position of the ith code component:

$$\begin{aligned} z_{ij} &= \frac{(I_{ij} + I_{i,j+1})}{\alpha} \\ &= R_{xi}(\tau - jt_0) + R_{xi}(\tau - (j+1)t_0) + n_{ij} \\ &= \bar{z}_{ij} + n_{ij} \end{aligned} \tag{13}$$

for an appropriately defined noise term n_{ij}. The signal portion of z_{ij}, denoted $\bar{z}_{ij}$ in the equation above (see Fig. 2),

$$\bar{z}_{ij} = \begin{cases} R_{xi} \times \left(1 + \frac{1}{N_i}\right)\left(\frac{\tau}{t_0} - j + 1\right) & \text{if } (j-1)t_0 \leqslant \tau < jt_0 \\ R_{xi} \times \left(1 + \frac{1}{N_i}\right) & \text{if } jt_0 \leqslant \tau < (j+1)t_0 \\ R_{xi} \times \left(1 + \frac{1}{N_i}\right)\left(j - \frac{\tau}{t_0}\right) & \text{if } (j+1)t_0 \leqslant \tau < (j+2)t_0 \\ 0 & \text{elsewhere} \end{cases} \tag{14}$$

If k is the correct range cell, i.e., $kt_0 \leqslant \tau < (k+1)t_0$, then the geometry of the correlation function (Fig. 1) leads to the estimator

$$\hat{\tau}_i - kt_0 = \frac{N_i I_{i,k+1} - I_{ik}}{(N_i - 1)(I_{ik} + I_{i,k+1})} = (\hat{\tau}_i \bmod t_0) \tag{15}$$

The right-hand side of this equation should be $(\hat{\tau}_i \bmod t_0)$, as already estimated accurately by clock-channel computations, within the expected noise deviation.

If there were no noise, the value of j yielding the maximum z_{ij} would be the correct range cell. But with noise, the cells on either side must also be scrutinized, as the spillover correlation and noise may cause us otherwise to choose the index of the wrong range cell.

Hence, let $\hat{k}_i$ be that j-index for which z_{ij} is maximum, and let k_i be $\hat{k}_i - 1$, $\hat{k}_i$, or $\hat{k}_i + 1$, whichever minimizes the difference between $\hat{\tau}_i$ and $\hat{\tau}_1$. This k_i is a high-likelihood range-cell estimate for the ith component because the maximum-likelihood range cell is certainly one of the three candidates, and the comparison of the symbol-fraction range offset, calculated as above, against the accurate determination of the clock

symbol fraction, rules out the other two candidates with high probability.

Analysis indicates that the probability of making the wrong choice would be very remote. As may be seen in the geometry of the cross-correlation function, shown in Fig. 1, the wrong pair of correlator values inserted into the symbol fraction formula above produces a significantly different estimate for $\hat{\tau}_i$ than for $\hat{\tau}_1$.

V. Range Calculation Algorithm

The procedure used by the $\nu\tau$ receiver to reconstruct the range is therefore basically the same as in the old-τ system, except for the way the component range cell determinations are made. Having read the 77 correlation values all at once after an integration interval T, the normal calculations on the two clock values determine the clock component range delay (mod $2t_0$) with high accuracy. The remaining 75 correlation values are grouped by component, and then a value $\hat{k}_i$ is chosen for each component. This $\hat{k}_i$ is the range-cell index j that maximizes the sum of adjacent correlation values $I_{i,j} + I_{i,j+1}$. Each of the three range-cell indices $\hat{k}_i - 1$, $\hat{k}_i$, and $\hat{k}_i + 1$ in turn is hypothesized to be the correct range-cell index. The appropriate correlator values for each of these indices are then used to compute a symbol-fraction offset using the estimator formula above. The candidate that comes closest to the clock-channel measurement wins.

Acknowledgments

The author gratefully acknowledges the encouragement and assistance of several colleagues. Paul Headley and Mike Rodriguez were instrumental in setting up the first meeting to discuss the possibilities of reconsidering multicomponent-uplink ranging. Shirley Peak and Bob Bunker performed a custom-VLSI feasibility study to determine whether conventional correlator channels could be built economically. John R. Smith performed a conceptual design of the VLSI correlators and spent a significant amount of time with the author and others discussing the signal detection and correlation method. Harold Baugh answered many questions concerning the current μ ranging system and its planned upgrade. John Smith also proposed the logic simplification discussed in the companion article [7].

References

[1] S. W. Golomb, *et al.*, *Digital Communications with Space Applications*, Englewood Cliffs, New Jersey: Prentice-Hall, 1964.

[2] J. H. Yuen (ed.), *Deep Space Telecommunications Systems Engineering*, JPL Publication 82-76, Jet Propulsion Laboratory, Pasadena, California, pp.123-178, July 1982.

[3] R. C. Tausworthe, "Optimal Ranging Codes," *Transactions of the IEEE PTG-SET*, vol. 10, no. 1, pp. 19-30, March 1964.

[4] R. C. Tausworthe, "Ranging the 1967 Mariner to Venus," in *Proc. of IEEE National Convention*, pp. 294-295, New York, March 1967.

[5] J. H. Yuen (ed.), *Deep Space Telecommunications Systems Engineering*, JPL Publication 82-76, Jet Propulsion Laboratory, Pasadena, California, pp. 160-161, July 1982.

[6] W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.

[7] R. C. Tausworthe and J. R. Smith, "A Simplified, General-Purpose Deep-Space Ranging Correlator Design," *TDA Progress Report 42-91*, vol. July-September, Jet Propulsion Laboratory, Pasadena, California, November 15, 1987.

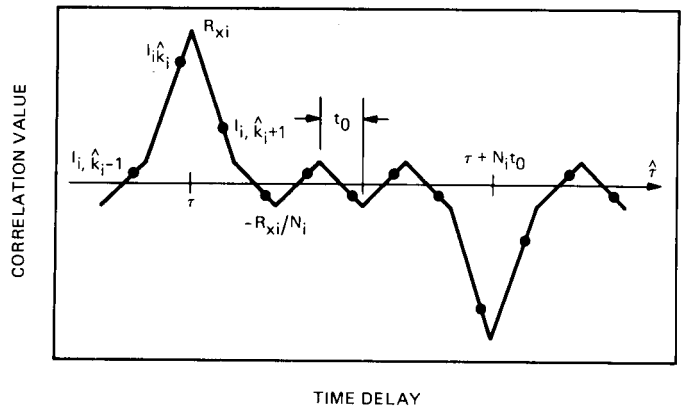


Fig. 1. Composite-code cross-correlation function $R_{xi}(\hat{\tau})$

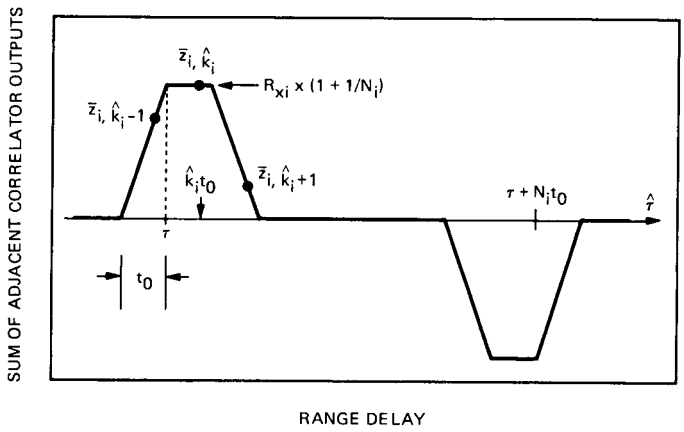


Fig. 2. Sum of adjacent correlator outputs versus range delay

Power Density Measurements in the Near Field of the DSS 13 26-Meter Antenna

E. B. Jackson
TDA Science Office

M. J. Klein
Space Physics and Astrophysics Section

Power density measurements were made at DSS 13 in the near field of the 26-m antenna to determine if radio frequency (rf) fields generated by the 20-kW transmitters could be responsible for the failure of three solid state (FET) rf amplifiers. These amplifiers are used in the SETI Radio Spectrum Surveillance System, which is currently located at the site. Measurements were made independently for one transmitter at 2115 MHz, one at 7150 MHz, and both transmitters together. Measurement results are tabulated and compared with predicted power densities under the measurement conditions. The results agree with predictions within a factor of two; the predictions appear to give "worst case" values. Measurements indicated that amplifier failures are not attributable to the transmitter.

I. Introduction

One of the field test activities for the Search for Extraterrestrial Intelligence (SETI) project at Goldstone involves the use of a Radio Spectrum Surveillance System (RSSS) to survey and characterize the radio frequency interference in the band from 1 to 10 GHz. Shortly after the initial survey was under way, three of the seven radio frequency (rf) amplifiers in the system failed at nearly the same time. Discussions with the manufacturer raised questions about potential damage from rf fields produced by the transmitters on the 26-m antenna, which is located a few hundred meters away. A plan to measure the rf fields at the SETI RSSS location was developed and carried out in July 1986.

Concern about mutual interference among co-located transmitting and receiving systems at the DSN complexes is clearly not unique to SETI.[1] The Planetary Radar Transmitter at DSS 14 is known to interfere with reception at DSS 15, which is located just 350 meters away. When transmitting near 8500 MHz, the radar interferes with VLBI data systems at the Mojave Base Station, which is located some 10 km away.

This article documents the measurements that were made and compares the results with expected levels based on calculations of rf field strengths in the near field of the 26-m antenna at DSS 13. Similar measurements were performed in 1970 on a 64-meter antenna [1].

[1] "RFI Analysis Final Report," ER 81-18, produced by Ford Aerospace under JPL Contract 956094, December 7, 1981.

II. Background

The SETI RSSS [2] is a scanning spectrometer with the capacity to step through a sequence of seven contiguous rf bands from 1 to 10 GHz. The system is equipped with a 0.91-meter-diameter parabolic antenna that can be stepped in azimuth under computer control. The elevation angle can be set manually and locked into position. The antenna is fed with a linearly polarized pyramidal log-periodic feed.

The system was installed in a van located 288 meters east of the DSS-13 26-m antenna (see Fig. 1). The antenna assembly, mounted on top of the van, is approximately 19 meters below the intersection of the azimuth and elevation axes of the 26-m antenna.

The 26-m antenna is equipped with transmitters at 2115 MHz and at 7150 MHz. Both transmitters, each with a nominal power output of 20 kW, can be radiating at the same time. A safety interlock system prevents transmitter operation when the antenna is tipped to elevation angles of less than 10 degrees.

As a result of altitude differences in the local terrain and the 10-degree elevation interlock, the rf axis of the main beam of the 26-m antenna, when transmitting, will always be at least 13 degrees above the RSSS antenna. Therefore, only off-axis responses of the 26-m antenna pattern would radiate power into the RSSS antenna. It is these power levels that were measured during the experiment.

III. Measurements

A Hewlett-Packard 436A Digital Power Meter was used to measure the detected power levels at the output of the RSSS antenna (Fig. 2). The power meter was equipped with a wideband detector (HP 8484A) which is sensitive from 10 MHz to 18 GHz. The RSSS antenna was pointed directly at the 26-m antenna, while the latter, within the elevation interlock limitation, was incrementally positioned to maximize the detected output of the power meter. The measurement procedure was independently performed for the two transmitters. The peak detected power levels and the corresponding 26-m antenna pointing angles were recorded.

The power density value at the input of the RSSS antenna can be derived from the detected power levels if the effective area of the receiving antenna is known. The effective area of an antenna, expressed in square meters, is given by the expression

$$A_e = \frac{g\lambda^2}{4\pi} \tag{1}$$

where g is the antenna gain and λ is the wavelength of the transmitted signal. The gain of the RSSS antenna is 23.4 dBi at 2115 MHz and 30.7 dBi at 7150 MHz. Power density S, in watts per square meter, is given by the expression

$$S = \frac{k(p)P}{A_e} \tag{2}$$

where $k(p)$ is the polarization factor, P is the measured power, and A_e is calculated from Eq. (1) for each wavelength. For this measurement, $k(p) = 2$ to account for the fact that the transmitter feeds are circularly polarized and the RSSS antenna is linearly polarized.

The results of the measurements and subsequent calculations are summarized in Table 1.

IV. Discussion

It is useful to compare the measured power densities with the predicted values that can be calculated for the near field of the 26-m antenna. D. A. Bathker (private communication) has estimated that the antenna gain, approximately 13 degrees off axis in the near field, should be about 3 dB above isotropic (dBi). For this estimate, the diffraction at the edge of the subreflector is estimated to be –6 dB. The on-axis gain of the common aperture (multi-frequency) feed horn is +22 dBi, and the illumination taper of the subreflector is –13 dB. Bathker estimates that the residual +3 dBi is distributed within the zone some 13 to 18 degrees off axis.

The power density $S(r)$, in watts per square meter, at distance r from the transmitting antenna is given by the expression

$$S(r) = \frac{P_t G}{4\pi r^2} \tag{3}$$

where $P(t)$ is the transmitter power (20 kW) and G is the numerical value of the off-axis antenna gain described above ($G = 2$ for the 3 dBi). According to Eq. (3), the power density for either frequency at the RSSS antenna, located in the near field at distance $r = 288$ meters, should be approximately 0.038 W/m^2. Note that this value is independent of frequency.

V. Conclusions

The measured values of the power density agree, within a factor of two, with the estimated value of 0.038 W/m^2. The measured values at both frequencies fall below the estimate, which can be considered to be a "worst case" prediction. This

result supports the validity of the power density estimates in the near field of the 26-m antenna at DSS 13.

The results of these measurements further suggest that the amplifier failures were not induced by the rf fields from the 26-m transmitters. Both measured and predicted power levels were far below the maximum in-band tolerance (+20 dBm) specified by the amplifier manufacturer. Subsequent tests of the FET amplifiers by the manufacturer confirmed this conclusion.

These results do suggest, however, that care should be taken if bipolar transistors are used in second or succeeding stages of rf amplification. Power feedthrough from first (and succeeding) FET stages could damage the bipolar transistors if the unit is subjected to rf fields similar to those measured in this study.

References

[1] D. A. Bathker, "Predicted and Measured Power Density Description of a Large Ground Microwave System," TM 33-433, Jet Propulsion Laboratory, Pasadena, California, April 15, 1971.

[2] B. Crow, A. Lokshin, M. Marina, and L. Chin, "SETI Radio Spectrum Surveillance System," *TDA Progress Report 42-82*, vol. April–June 1985, Jet Propulsion Laboratory, Pasadena, California, pp. 173–184, August 15, 1985.

Table 1. Results of power density measurements at DSS 13

Parameters	2115-MHz transmitter 20 kW	7150-MHz transmitter 20 kW	Both transmitters on
Maximum detected power	+5.0 dBm 3.16 mW	+4.1 dBm 2.57 mW	+9.2 dBm 8.32 mW
26-m antenna position			
Azimuth angle	92.176 deg	92.620 deg	92.343 deg
Elevation angle	11.502 deg	11.547 deg	11.547 deg
RSSS antenna effective area	0.350 m^2	0.165 m^2	
Power density input to RSSS antenna	0.018 W/m^2	0.031 W/m^2	

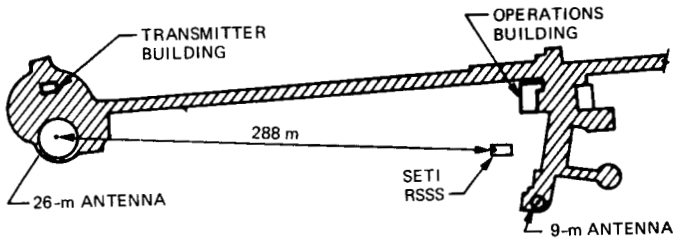


Fig. 1. Location of SETI RSSS on the Venus station

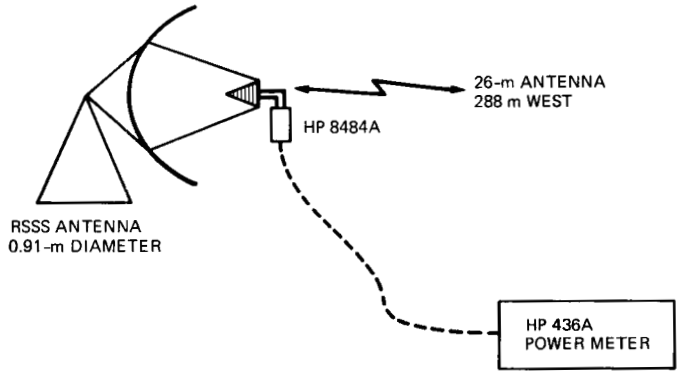


Fig. 2. Power density measurements for the 26-meter antenna